UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE ESTUDIOS ESTADÍSTICOS

TESIS DOCTORAL

Algunas aportaciones a la toma de decisiones en clasificación supervisada. Un enfoque bipolar

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Guillermo Villarino Gutiérrez

Directores

Daniel Gómez González
Juan Tinguaro Rodríguez
Rosario Cintas del Río

Madrid

Algunas aportaciones a la toma de decisiones en clasificación supervisada. Un enfoque bipolar

Memoria para optar al grado de doctor presentada por

GUILLERMO VILLARINO MARTÍNEZ

Facultad de Estudios Estadísticos

UNIVERSIDAD COMPLUTENSE

Tesis dirigida por

DANIEL GÓMEZ GONZÁLEZ, JUAN TINGUARO RODRÍGUEZ Y ROSARIO CINTAS DEL RÍO

15 DE JULIO DE 2019

UNIVERSIDAD COMPLUTENSE MADRID

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

D./Dña. GUILLERMO VILLARINO MARTÍNEZ,
estudiante en el Programa de Doctorado ANÁLISIS DE DATOS (DATA SCIENCE),
de la Facultad de Estudios Estadísticos de la Universidad Complutense de Madrid, como autor/a de la tesis presentada para la obtención del título de Doctor y titulada:

ALGUNAS APORTACIONES A LA TOMA DE DECISIONES EN CLASIFICACIÓN SUPERVISADA. UN ENFOQUE BIPOLAR.

y dirigida por: DANIEL GÓMEZ GONZÁLEZ, JUAN TINGUARO RODRÍGUEZ Y ROSARIO CINTAS DEL RÍO

DECLARO QUE:

La tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la Ley de Propiedad Intelectual (R.D. legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, modificado por la Ley 2/2019, de 1 de marzo, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita.

Del mismo modo, asumo frente a la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad del contenido de la tesis presentada de conformidad con el ordenamiento jurídico vigente.

En Madrid, a 15 de julio de 2019

VILLARINO MARTINEZ GUILLERMO - 50759730H

Firmado digitalmente por VILLARINO MARTINEZ GUILLERMO - 50759730H

Fdo.: _____

Esta DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD debe ser insertada en la primera página de la tesis presentada para la obtención del título de Doctor.

Algunas aportaciones a la toma de decisiones en clasificación supervisada. Un enfoque bipolar.

MEMORIA PRESENTADA PARA OPTAR AL GRADO DE DOCTOR POR

Guillermo Villarino Martínez

Departamento de Estadística y Ciencia de Datos
Facultad de Estudios Estadísticos
Universidad Complutense de Madrid

Julio 2019

A nuestro sabio despistado

Agradecimientos

No hay grandeza donde faltan la sencillez, la bondad y la verdad.

Leon Tolstoi

Esta memoria nace con la pretensión de arrojar algo de luz sobre el proceso de investigación llevado a cabo en los últimos años, desde los inicios y motivaciones del autor para desarrollar las ideas presentadas, pasando por la descripción de las distintas vías exploradas hasta los últimos resultados obtenidos en este campo y las aportaciones más relevantes.

Para bien o para mal, la existencia de este trabajo esta sin duda supeditada al acontecimiento más triste de mi vida, el fallecimiento del *Sabio Despistado*. Es entonces cuando mi planteamiento sobre la vida en términos generales sufre un giro importante, dando mayor valor al conocimiento y la *sabiduría* e incluso a aquella *cultura del esfuerzo* de la que tanto había oído hablar durante mi infancia y juventud. Por este motivo, el sentimiento de estas palabras transciende con creces el agradecimiento, entrando más bien en el terreno de la admiración. Admiración por el hombre, por el padre, por el profesional. Admiración por esa mente inquieta que no paraba de cavilar, por esas noches enteras en la cocina resolviendo problemas matemáticos por pura *pasión* por el conocimiento. Sin esa semilla, es evidente que mi camino habría sido bien diferente. Siento este trabajo como mi pequeño y humilde homenaje a la persona que más añoro en esta vida.

Dicho esto, es para mí muy importante agradecerle a la vida el hecho de haber *caído* en la Facultad de Estudios Estadísticos ya que, desde el minuto uno me he sentido como en mi propia casa. Me considero verdaderamente afortunado por haber tenido la oportunidad de estudiar y formarme en este centro y haber sido espectador y, en ocasiones, participe de los grandes cambios sufridos en los últimos años, desde el paso de Escuela Universitaria a Facultad, la creación los dos títulos de máster y, en especial, de este programa de doctorado en Análisis de Datos o *Data Science* del cual tengo el honor de haber sido el primer estudiante becado a tiempo completo. Mi agradecimiento a todas las personas de este centro, no solo personal docente e investigador sino también personal de administración y servicios y colegas

de doctorado con quienes he compartido muchos ratos de cafés y *chascarrillos* que han hecho más llevaderos los momentos de soledad propios del investigador novel.

Me gustaría resaltar que este trabajo no hubiera sido posible sin el incansable apoyo de mis supervisores Daniel Gómez González y J. Tinguaro Rodríguez así como de mi supervisora Rosario Cintas de Río, quienes han sabido guiar diligentemente mis pasos en este camino investigador, haciendo que aquellos momentos de mayor incertidumbre y dificultad resultaran más llevaderos. Así mismo, profundo agradecimiento a mi familia, mi adorada madre Mª Jesús, mis queridas hermanas Esther, Carol y Almu y las últimas incorporaciones a este equipo, mis sobrinas Paula, Iria y Alba y el casi recién llegado gran varón, Simón. Son estas pequeñas personas quienes dibujan la sonrisa más sincera en mi rostro.

Por último y, como suele decirse, en absoluto menos importante, agradecimientos a mis amigos y amigas, principales responsables de soportar mis posibles desmanes fruto de las épocas de intenso trabajo. Especial mención merecen mis *Hakuna*, mi familia *postiza* desde que elegí ser feliz hará ya cosa de ocho años, sin vuestra inyección de energía en cada ensayo, cada bolo y cada rato compartido, mi camino hubiera tenido una pendiente mucho mayor.

Resumen

Algunas aportaciones a la toma de decisiones en clasificación supervisada. Un enfoque bipolar.

Introducción

Las aplicaciones de la Ciencia de Datos y la Inteligencia Artificial adquieren cada vez mayor relevancia en el mundo actual. Como muchas de estas aplicaciones influyen en los procesos de toma de decisiones, en los que los decisores humanos asumen la responsabilidad de la decisión final y sus consecuencias, también existe una creciente necesidad de hacer de los seres humanos el centro del análisis de los datos. En este sentido, los responsables de la toma de decisiones deben tener capacidad de interpretar los modelos de aprendizaje automático que deben usarse para ayudar a tomar decisiones de gran importancia, ya que deben comprender y poder explicar por qué tomaron las decisiones que tomaron. De manera similar, en este esfuerzo por adaptarse más a las características y la naturaleza de los humanos, los métodos de la ciencia de la información también deben tener en cuenta: i) que los humanos se comunican de manera eficiente por medio del lenguaje natural, en el que los conceptos generalmente poseen una naturaleza imprecisa e incierta; y ii) que el razonamiento humano y la toma de decisiones típicamente se basan en un balance de informaciones de carácter positivo y negativo, es decir, tienen una naturaleza dicotómica o bipolar.

Objetivos y resultados

En esta memoria se exploran diversos métodos para la incorporación de un marco de *bipolaridad* en algoritmos de clasificación supervisada de naturaleza *soft*, entendiendo esta clase de algoritmos como aquellos que devuelven una puntuación en la etapa previa a la toma final de decisiones, en contraposición a los algoritmos de naturaleza puramente nítida, en los que se asigna a cada observación una clase predefinida. La incorporación de este marco bipolar, permiten explotar la información de carácter *soft* de un clasifica-

dor, evitando así, la utilización de reglas convencionales como la *regla del máximo*, y dotando al clasificador de una mayor flexibilidad al considerar distintas evidencias sobre la pertenencia de un objeto a cierta clase.

Se proponen en este trabajo dos grandes grupos de aproximaciones a *Nivel Global*, esto es, susceptibles de aplicación en la etapa final de toma de decisiones de cualquier clasificador, para abordar este problema. Por un lado las que están orientadas a la consideración de clasificadores probabilísticos, esto es, cuya evidencia viene dada en forma de probabilidad. Por otro, aquellas cuyo objetivo es aprovechar la información dada por clasificadores de tipo *fuzzy* o difuso, dada en este caso por grados de pertenencia difusos (que miden el grado de verdad de la afirmación $x \in C_j$, siendo x una observación del conjunto de datos y $C_j, j = 1, ..c$ una de las c clases del problema de clasificación) con distintas propiedades que las probabilidades. Se destaca que estos métodos a *Nivel Global* son susceptibles de aplicación a cualquier algoritmo de clasificación de naturaleza *soft*.

En el contexto de los algoritmos difusos y, concretamente sobre los Sistemas de Clasificación Basados en Reglas Difusas (SCBRDs), se desarrolla un paradigma bipolar a *Nivel Local* o de reglas, que permite representar el conocimiento presente en la base de reglas de la clasificación. Por tanto, es este un método que, por ahora, se encuentra restringido al marco de los SCBRDs teniendo, no obstante, gran proyección debido a sus buenas propiedades.

Un completo marco teórico de los diferentes enfoques bipolares considerados se desarrolla aquí con el objetivo de proporcionar un esquema matemático flexible, susceptible de aplicación bajo distintas configuraciones paramétricas e incluso modificaciones en pos de la adaptación a un contexto determinado. Además, algunas configuraciones paramétricas se evalúan a través de estudios experimentales rigurosos en conjuntos de datos de la vida real para evaluar el comportamiento del esquema bipolar en el contexto de la clasificación supervisada.

Conclusiones

Como conclusión general de este trabajo, la consideración de un marco bipolar en el contexto de los algoritmos de clasificación en *Data Science* constituye una solución adecuada para hacer frente a las tareas de asignación de clases, lo que permite al clasificador aprovechar una estructura de evidencia pareada y, por lo tanto, otorgar al clasificador un proceso de toma de decisiones con mayor flexibilidad. En cuanto al paradigma a *Nivel local* propuesto en el marco de los SCBRDs, se evidencia una sinergia positiva entre estos dos métodos de representación de conocimiento basados en el ser humano, abriendo así una línea interesante de investigación futura.

Abstract

Some contributions to decision making in supervised classification. A bipolar approach.

Introduction

Data Science and artificial intelligence applications have become more and more relevant in today's world. As many of these applications influence decision-making processes, in which human decision-makers take the responsibility of the final decision and its consequences, there is also an increasing need for making humans the centre of the data analysis itself. For instance, machine learning models used to aid high stake decision-making have to be interpretable by these decision makers, as they need to understand and be able to explain why they made the decisions that they made. Similarly, in this effort to become more adapted to human characteristics and nature, data science methods should also take into account i) that humans best communicate by means of natural languages, in which concepts usually possess an imprecise, uncertain nature; and ii) that human reasoning and decision-making typically proceeds by weighing information pieces with a positive or negative character, i.e. they have a dichotomous or bipolar nature.

Objectives and outcomes

In this human-inspired context, the here presented report is focused on exploring some new methods for the application of a *bipolar* framework upon *soft* supervised classification algorithms. A *soft* classifier reaches a numeric score prior to the final decision-making process. Some useful ideas to exploit this *soft* information are proposed, thus avoiding the "blind" use of the so called *maximum rule* and giving the classifier a greater flexibility by considering different kinds of evidence about the belonging of an object to a certain class.

Two large groups of approaches on a *Global Level* are proposed in this work. On the one hand, several methods orientated to the consideration of

probabilistic classifiers, that is, the evidence given in the form of a probability. On the other hand, those with the aim of taking advantage of the information given by *fuzzy* classifiers, considering in this case *fuzzy* membership degrees that have different properties to those of the probabilities. It should be noted that all these *Global Level* methods are susceptible to be applied upon any *soft* classification algorithm.

In the context of fuzzy algorithms and, specifically, considering the Fuzzy Rule Based Classification Systems (FRBCSs), a bipolar paradigm on a *Local Level* is developed. The here denoted *Local Level* allows for the knowledge contained in the rule base of the classifier to be represented. Therefore, this method, for now, is restricted to the FRBCS framework, having, however, great projection due to its good properties.

A complete theoretical framework, regarding the different bipolar approaches considered, is developed here with the aim of giving a flexible mathematical scheme, which can be used in several different parametric configurations and even modified to adapt it to a certain context. Moreover, some parametric configurations are evaluated through rigorous experimental studies on real life datasets in order to assess the behaviour of the bipolar scheme in the context of supervised classification.

Conclusions

As a general conclusion of this work, the consideration of a bipolar framework in the context of classification algorithms in Data Science stands a suitable solution to deal with class assigning tasks, allowing the classifier to take advantage of a paired evidence structure and therefore endowing the classifier with a more flexible decision making process. Regarding the *Local level* paradigm in FRBCSs, a positive synergy between these two human-based knowledge representation methods is evidenced, opening in this way an interesting line of future research.

Índice

Índice de figuras

Índice de Tablas

Capítulo 1

Introducción

Todo tiene su final..
Héctor Lavoe

Resumen: En este capítulo introductorio, se detallan las motivaciones del autor para la realización de este trabajo de investigación (Sección 1.1) así como los objetivos planteados (Sección 1.2) y las principales contribuciones realizadas (Sección 1.3).

1.1. Motivación

Este trabajo se inspira en la propensión humana a la duda frente a sentencias categóricas al respecto de una decisión concreta. Es bien cierto que, como seres humanos, el hecho de confiar en las decisiones no suficientemente fundamentadas se considera, al menos, arriesgado. Pues bien, en el contexto de la Inteligencia Artificial (IA), son muchas las decisiones y sentencias categóricas realizadas ya no por personas sino por máquinas. No es de extrañar, por tanto, que exista cierta inquietud social al respecto.

Cuando de tomar decisiones se trata, resulta necesario disponer de la mayor cantidad posible de información al respecto de cara a valorar las consecuencias de cada una de las alternativas. Así, se puede pensar en el caso de la predicción meteorológica. Es habitual, hoy en día, recibir esta información en forma de puntuaciones de naturaleza *soft*, en concreto, probabilidades, por lo que es el propio usuario quien toma la decisión final sobre la pregunta de naturaleza nítida *¿lloverá mañana?*. De lo contrario, en un contexto dicotómico, un modelo de predicción puede verse obligado a tomar una decisión nítida (si/no) y, por tanto, a aplicar sobre las citadas puntuaciones una regla de decisión, siendo capaz así de dar respuesta a la pregunta planteada, imaginemos que es un *si*. Todo usuario precavido sacará mañana su

paraguas. Ahora bien, esa respuesta nítida bien puede haber sido extraída de unas probabilidades estimadas por el modelo como por ejemplo $p_{si} = 0.51$ y $p_{no} = 0.49$. Bajo esta premisa, las evidencias obtenidas sobre la probabilidad de ambos eventos tienen valores cercanos y, por ello, la decisión está basada de alguna forma en información confusa. Posiblemente mañana los paraguas permanezcan secos. Esta situación ilustra la necesidad de considerar la información *soft* de los clasificadores, dotándola de un carácter más amplio.

Usualmente, el mecanismo de toma de decisiones se basa en un balance entre los aspectos positivos y negativos que presenta cada una de las opciones, la famosa lista de *pros* y *contras*. Inspirado en esta característica humana se desarrolla la idea de *bipolaridad*, que trata de representar estos afectos de naturaleza contrapuesta (pros y contras) mediante formalismos matemáticos en el contexto de la representación del conocimiento. Parece pues razonable explorar la capacidad de toma de decisiones de este paradigma. Este trabajo se enmarca en lo que se conoce como ciencia de datos (*Data Science*) cuyo objetivo es la extracción del conocimiento subyacente a los datos.

En la ciencia de datos, uno de los temas más importantes es la clasificación, y particularmente las tareas de clasificación supervisadas. En la literatura, existe una gran diversidad de algoritmos de clasificación supervisados, enfoques y aplicaciones, según las tareas específicas, el tipo de datos, las características o la eficiencia [49, 51]. Normalmente, en un contexto de clasificación supervisada, el objetivo principal es poder clasificar un conjunto de elementos en clases en función de una muestra de entrenamiento o conjunto de datos que proporciona ejemplos de asociación entre elementos y clases, y que se utiliza para entrenar los clasificadores con el fin de generalizar adecuadamente las asociaciones observadas, es decir, ajustar los modelos de clasificación a los datos observados.

Tras esta fase de entrenamiento, los algoritmos de clasificación proporcionan un mecanismo que permite clasificar nuevos elementos (nuevas consultas o elementos en una muestra de prueba) en algunas de las clases conocidas. En esta tarea de clasificación, es importante diferenciar entre el proceso de aprendizaje interno, que con frecuencia asigna a cada pareja (elemento, clase) una puntuación de tipo *soft* (por ejemplo, una probabilidad, un grado difuso, una posibilidad) que evalúa la fuerza de la asociación de cada elemento con cada clase, y el mecanismo o etapa de decisión final que rige cómo se asignan los elementos a una sola clase en función de estas puntuaciones.

En este sentido, la mayoría de los clasificadores solo asignan cada elemento a la clase que obtuvo el *soft score* más alto, un procedimiento de decisión que generalmente se conoce como *regla del máximo*. No obstante, a pesar de su amplia aplicación, esta regla puede no ser el procedimiento de decisión óptimo en todos los contextos. Por ejemplo, esta regla puede no necesariamente aprovechar toda la información relevante contenida en

las puntuaciones, particularmente en contextos en los que las clases presentan algún tipo de correlación o estructura de interdependencia. Una de las principales motivaciones de este trabajo es el estudio de este problema para mejorar la adaptación de los clasificadores a cada contexto de aplicación específico y, por lo tanto, también su rendimiento.

Por otro lado, no es desdeñable el interés despertado por los métodos de representación de información mediante reglas lingüísticas. Es este un esquema de tratamiento de datos basado en el lenguaje y, por tanto, dotado de una indudable interpretabilidad. Es esta característica la que hace de este tipo de paradigmas de representación del conocimiento una interesante alternativa en el contexto de la ciencia de datos y la clasificación supervisada. Más allá, la posible sinergia entre este esquema y la idea de *bipolaridad* merece ser evaluada.

1.2. Objetivos

En este trabajo se incide en la necesidad de conocer y, en su caso, manejar la información de carácter numérico (puntuación) alcanzada por la mayor parte de los algoritmos de clasificación supervisada, evitando de esta forma la aplicación automática de la regla del máximo sobre el vector de puntuaciones finales. Así mismo, se asume la existencia de una inherente relación entre las distintas clases objeto de la clasificación a través de una estructura de *disimilitud* subyacente en los datos. Mas allá, se propone la evaluación de posibles sinergias entre dos de las filosofías de representación formal del conocimiento inspiradas en el ser humano, como *lógica difusa* y *bipolaridad*.

A continuación se concretan los objetivos generales planteados en esta memoria y los objetivos específicos asociados a cada uno de ellos.

Objetivo 1: Explotar la información de naturaleza *soft* dada por cualquier algoritmo de clasificación supervisada en la etapa previa a la toma de decisiones, definiendo un nuevo marco para la toma de decisiones basado en la Representación Bipolar del Conocimiento (RBC).

1. Definir formalmente el concepto de estructura de disimilitud entre grupos de instancias pertenecientes a distintas clases.

2. Estudiar las relaciones existentes entre estructuras de disimilitud bajo distintos tipos de agregaciones.

3. Plantear un sistema de toma de decisiones global *a posteriori* basado en información bipolar aplicable a cualquier tipo de algoritmo de clasificación probabilístico o difuso.

4. Proponer un paradigma de clasificación bipolar robusta basado en re-

plicación y agregación difusa en el contexto de la clasificación probabilística.

Objetivo 2: Proponer una extensión del marco general de la inferencia en Sistemas de Clasificación Basados en Reglas Difusas (SCBRDs) basada en Representación Bipolar del Conocimiento (RBC).

1. Extender el sistema de ayuda a la decisión global al contexto de los sistemas de clasificación basados en reglas (SCBRDs).
2. Proponer un nuevo mecanismo de representación bipolar del conocimiento a nivel local, esto es, a nivel de reglas en el marco de los SCBRDs.
3. Extender el Método de Razonamiento Difuso (MRD) de los SCBRDs para el manejo de información de carácter bipolar

Objetivo 3: Casos prácticos: por una parte, se pretende aplicar algunos de los avances al estudio de datos de siniestralidad vial proporcionados por la Dirección General de Tráfico (DGT), por otra, se considera la representación de la información *soft* en el caso especial de clasificación de segmentos en el contexto de la detección de bordes en imágenes.

1.3. Contribuciones

En esta sección se enumeran las contribuciones extraídas del trabajo presentado en esta memoria, en un orden cronológico:

Contribución 1: Villarino, G., Gómez, D., Cintas, R., Rodriguez, J.T. (2016): *Metodología de minería de datos para el estudio de tablas de siniestralidad vial.* In Proceedings of CAEPIA-STYLF Congress 2016 599–608.

Contribución 2: Villarino, G., Gómez, D., Rodríguez, J. T. (2017): *Improving Supervised Classification Algorithms by a Bipolar Knowledge Representation.* Advances in Intelligent Systems and computing **643** 518–529. doi:10.1007/978-3-319-66827-7_48.

Contribución 3: Flores-Vidal, P. A., Gómez, D., Montero, J., Villarino, G. (2017): *Classifying segments in edge detection problems.* 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Nanjing, 2017, 1–6. doi:10.1109/ISKE.2017.8258764

Contibución 4: Villarino, G., Gómez, D., Rodriguez, J.T. et al. (2018): *A bipolar knowledge representation model to improve supervisedfuzzy classification algorithms.* Soft Compututing **22** 5121–5144.

doi:10.1007/s00500-018-3320-9.

Contribución 5: Villarino, G., Gómez, D., Rodríguez, J.T. (2018): *Assessing the performance of bipolar classifiers in three-class problems.* In Proceedings of CAEPIA Congress 2018 306–314

Contribution 6: Flores-Vidal, P. A., Villarino, G., Gómez, D., Montero, J.(2019): *Classifying segments in edge detection problems.* International Journal of Computational Intelligence Systems **12** (1) 367–378.
doi:10.2991/ijcis.2019.125905653

Contribución 7: Villarino, G., Gómez, D., Rodríguez, J.T., Fernández, A. (2019): *A new exploitation scheme in the context of bipolar classifiers.* ESCIM 2019 Congress. In press.

Contribución 8: Villarino, G., Rodríguez, J.T., Gómez, D., Fernández, A. (2019): *Extending the Fuzzy Inference by Bipolar Representation in a Classification Context: New Methods and Models.* [En preparación]

Capítulo 2

Preliminares

Dadme un punto de apoyo y moveré el mundo.

Arquímedes de Siracusa

Resumen: En este capítulo se repasan algunos de los conceptos preliminares que constituyen el punto de apoyo necesario para la elaboración de esta memoria. En primer lugar, se dedica la Sección 2.1 a contextualizar el problema raíz de esta investigación, el problema de la siniestralidad vial. A continuación, en la Sección 2.2 se presentan los conjuntos difusos como herramienta para el manejo de datos mediante etiquetas lingüísticas y en la Sección 2.3 se detallan los conceptos básicos manejados en la Representación Bipolar del Conocimiento (RBC). En las Secciones 2.4 y 2.5 se repasan respectivamente, las definiciones de distintos tipos de clasificadores y algunos de los principales algoritmos de clasificación supervisada en el contexto de la *Minería de Datos* (MDD), distinguiendo entre las dos familias de clasificadores considerados en este trabajo en función del tipo de puntuación alcanzada como probabilísticos y difusos. En la Sección 2.7 se da una visión global sobre las medidas de evaluación de la eficacia predictiva de los clasificadores en el marco de la MDD. Finalmente, la Sección 2.8 contiene las nociones sobre metaheurísticas de búsqueda en el contexto de la clasificación supervisada.

2.1. Siniestralidad vial. DGT

La accidentalidad en las carreteras ha sido, desde la generalización del uso de vehículos a motor, una de las principales causas de muerte en España y por

ello es foco de gran preocupación para la sociedad y sus autoridades. Han pasado muchos años desde el máximo histórico de fallecidos en accidentes de tráfico en España de 1989. Aquel aciago año, último de la década de los ochenta y en pleno aumento del parque de vehículos automóviles, 9.344 personas perdieron su vida en un accidente de tráfico. Entonces no se llegaba a 15 millones de vehículos en total.

Debido a los esfuerzos realizados y a la gran cantidad de recursos destinados, estas elevadas cifras han experimentado, afortunadamente, un descenso muy significativo. Entre las causas de esta disminución se encuentran la mayor concienciación de la sociedad en materia vial, la mejora de las infraestructuras de la red de transportes, los cambios legislativos, los grandes avances en materia de seguridad de los vehículos automóviles y de detección de infracciones mediante cinemómetros y cámaras de seguridad.

España se encuentra entre los 10 países con menor siniestralidad de la Unión Europea, y en concreto ocupa la 7ª posición. La sociedad debe ser consciente de que reducir las cifras actuales no es una tarea sencilla y que, para ello es necesario impulsar políticas eficaces de seguridad vial, basadas en la evidencia científica; políticas que tengan en consideración los diferentes sectores implicados y que comprometan de manera efectiva a estos sectores, tanto públicos como privados, en la reducción de las cifras de siniestralidad vial. Las 1903 víctimas mortales y los 10444 heridos graves, según los informes policiales, ocasionados en vías urbanas e interurbanas [1] son motivación más que suficiente para mejorar las estrategias que nos lleven a erradicar el grave problema de salud que suponen las lesiones por accidente de tráfico.

Con el principal objetivo de tratar de analizar la gravedad de los accidentes de tráfico y entender sus principales causas, en [79] se presenta una metodología para el estudio de datos de siniestralidad vial en España. Es este un problema ya profundamente estudiado en anteriores trabajos del autor, por ello, es más que conocido su carácter especialmente complejo en lo que se refiere a la clasificación. Así, afortunadamente en términos sociales pero desgraciadamente desde el punto de vista de los paradigmas de clasificación, se trata de un problema de clasificación de datos con clases poco representadas (clases desbalanceadas o con baja incidencia del suceso de interés) hasta el punto de presentar la irrisoria tasa de evento (en este caso accidentes con resultado mortal) del 1.2 % en la población general de accidentes, reduciendo su valor en caso de la subpoblación de turismos hasta el 0.64 %.

Este problema especial de clasificación se considera la semilla inicial del desarrollo de las líneas de investigación seguidas en este trabajo, siempre orientadas al tratamiento de la información de naturaleza *soft* extraída de la aplicación de clasificadores a los datos, trascendiendo así la peligrosa consideración de la regla del máximo como criterio único de cara a la asignación

[1] Las principales cifras de la siniestralidad vial en España 2012. Dirección General de Tráfico (DGT)

nítida a una u otra clase.

.

2.2. Lógica y Conjuntos Difusos

Si por algo se caracteriza el ser humano es por su capacidad de generar patrones relativamente complejos de comunicación que engloban simultáneamente conceptos extremadamente concretos y entidades abstractas e imprecisas. Esta capacidad para razonar en base a la imprecisión aporta una gran flexibilidad para el entendimiento del lenguaje en términos subjetivos. A nivel matemático, una de las herramientas que permiten este tratamiento formal es la lógica fuzzy o borrosa o difusa [87], la cual, a partir de una serie de trabajos realizados en la segunda mitad del siglo pasado, ha posibilitado el desarrollo de modelos para un tratamiento sistemático del lenguaje que permita, al menos, una implementación informática progresivamente más sofisticada y potente de conceptos y razonamientos de corte lingüístico, en lo que se ha venido conociendo sucesivamente como razonamiento aproximado (*approximate reasoning* [40]), computación suave (*soft computing* [91]) o computación con palabras (*computing with words* [92]. Las líneas de investigación que conforman estas disciplinas han tenido, desde su origen, un gran impacto en el progreso posterior de la inteligencia artificial, siendo más que notable su influencia en el campo del aprendizaje automático [3, 23, 44, 50, 73].

Como consecuencia de ciertas reflexiones y propuestas de mejora realizadas en torno a esas ideas fundamentales, aparece la noción de variable lingüística [90], quizás una de las más afortunadas y con más impacto práctico sobre los campos de investigación antes referidos. Gracias a este concepto es posible modelar con cierto rigor realidades fuertemente ligadas a la operación con lenguajes naturales, que de otro modo hubiera sido difícil tratar desde una perspectiva matemática formal. Es importante recordar y destacar aquí que buena parte del conocimiento, muchas veces intuitivo y no formalizado, que manejan los expertos de diversas materias (por ejemplo, la medicina, la prospección geológica y, como no, la gestión de desastres y emergencias) suele ser expresado mediante expresiones del lenguaje común, y la facilidad que estas herramientas formales proveen para su tratamiento informático y automatizado constituye una de las bases para el éxito de muchos desarrollos que se han llevado a cabo en el campo de la inteligencia artificial a partir de mediados de la década de los años 70 del siglo pasado.

Formalmente, la lógica borrosa se presenta como una generalización de la lógica clásica binaria, en el que el conjunto de valores de verdad $\{0, 1\}$ es reemplazado por el intervalo $[0, 1]$ o, de manera más general, por un retículo ordenado $L = (L, \leq_L, \vee, \wedge, 0, 1)$ con elementos mínimo (0) y máximo (1). De este modo, dado un predicado P y un universo de discurso X, que contiene los objetos sobre los que ese predicado actúa, se evalúa la veracidad de la

afirmación "x es P", $x \in X$, asignándole un valor de verdad en L. En otras palabras, se define el conjunto borroso P, que da extensión al predicado P, asignando a cada elemento $x \in X$ el valor de verdad de la afirmación "x es P", lo cual se realiza a través de la función de pertenencia $\mu_p : X \to L$ de ese conjunto borroso.

Como generalización de la lógica clásica, los operadores borrosos para la negación, conjunción y disyunción han de generalizar a los clásicos. Básicamente, esto implica imponer condiciones de contorno para dichos operadores sobre los elementos extremos 0 y 1 del retículo L. Así, por ejemplo, un operador difuso $n : L \to L$ que represente la negación ha de cumplir $n(0) = 1, n(1) = 0$, y de igual modo un operador $T : L \times L \to L$ para la conjunción ha de satisfacer $T(l, 1) = T(1, l) = l, \forall l \in L$. Obviamente, en un contexto borroso puede existir un gran número de operadores de negación, conjunción y disyunción, y esta es otra de las características de las lógicas borrosas: la generalización de los operadores lógicos tradicionales no está (en absoluto) unívocamente determinada. En este sentido, es más habitual hablar de tipos y familias de operadores para la conjunción y la disyunción, siendo el tipo más común el dado por t-normas y t-conormas, aunque también es posible encontrar uninormas (vease [63]), operadores recursivos (ver [17]) o funciones de agregación (ver [37, 38, 56]), entre otros.

Especial atención merece el concepto de variable lingüística [89, 89, 90] por su capacidad de manejar matemáticamente conceptos imprecisos con fronteras no nítidas entre categorías. En muchas disciplinas que tienen por objeto de estudio el ser humano existe una clara necesidad de tratar con predicados de tipo difuso en tanto los juicios proporcionados por las personas son, a menudo expuestos en estos términos. Para aclarar estas ideas, es conveniente mostrar con un ejemplo la distinción entre variable lingüística y variable nítida.

Diferencia entre variable lingüística y variable nítida. Con el objetivo de relatar las diferencias existentes entre una variable nítida (típicamente una variable de naturaleza categórica o nominal) y una variable difusa/lingüística con el mismo número de categorías, se puede pensar en la medición de una característica como la altura de una persona en términos categóricos. A nivel variable nítida, se trata de una tramificación de una variable continua en origen y con rango total $R = (R_i, R_s)$, mediante la cual se agrupan los valores comprendidos en un rango determinado $L_j = (L_i, L_s) \subset R$ asignando una etiqueta de valor C_j para todo el conjunto de instancias pertenecientes a este rango L_j con independencia de su posición dentro de dicho intervalo. A nivel conceptual, esta simplificación puede resultar poco realista ya que puede suceder que dos elementos cercanos a los extremos superior e inferior de dos intervalos adyacentes tengan una mayor similitud que dos elementos situados a gran distancia dentro del mismo intervalo. Por esta razón, la apa-

rición del concepto de variable difusa y su extensión a nivel semántico, la variable lingüística, hace posible un tratamiento más flexible relajando las fronteras entre categorías adyacentes. La diferencia fundamental entre ambos tipos de variables, que se deriva de las propiedades de los conjuntos difusos y nítidos, es la propiedad de pertenencia simultánea a más de una categoría. Esta propiedad supone que un elemento puede tener cierto grado de compatibilidad con un conjunto de etiquetas lingüísticas que se encuentran de alguna forma solapadas en el espacio continuo. Ilustremos gráficamente esta diferencia existente entre ambos tipos de variables. Para ello, en la Figura 2.1 se muestra la representación del concepto *altura de una persona* desde las dos perspectivas a comparar. Por un lado, en la figura de la izquierda, se tiene la tramificación clásica de esta variable de naturaleza continua en 3 intervalos de altura (Baja, Media, Alta) en la que es clara la separación entre cada una de las categorías de altura definidas, de manera que un individuo ha de enmarcarse en una y solo una de ellas. No obstante, se puede plantear una tramificación de esta variable continua de forma que se permita el solapamiento natural entre las categorías adyacentes soslayando la limitación de pertenencia a un único grupo que presentan las variables de tipo categórico (ordinal en este caso).

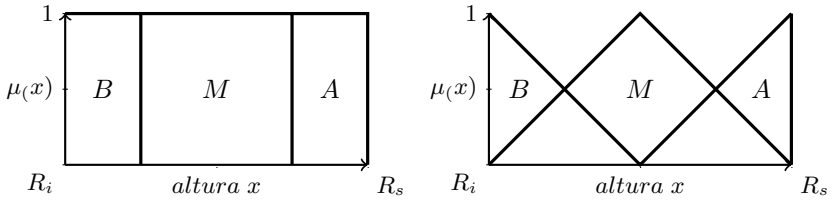


Figura 2.1: Representación del concepto *altura de una persona* mediante una variable categórica-ordinal clásica (izquierda) y una variable difusa lingüística con función de pertenencia triangular (derecha).

En la ilustración de la derecha de la Figura 2.1 se representa una posible variable lingüística asociada al concepto altura de una persona, en la que considerando igualmente 3 etiquetas o categorías, se modela la pertenencia a cada una de ellas como una función *triangular* que alcanza su máximo en el entorno del centro, o extremos superior o inferior, de los intervalos que forman cada categoría según la posición de la misma, siendo menor su valor a medida que nos alejamos de estos puntos en los intervalos. Cabe destacar que existe un claro solapamiento de las funciones triangulares que definen las clases a lo largo de toda la longitud de los intervalos, de esta forma un individuo que no se encuentre en el comentado centro o extremo del intervalo según el caso (donde los grados de pertenencia son igual a la unidad), presentará un grado de pertenencia no nulo a cada una de las categorías adyacentes, evaluando para definir la pertenencia final, el valor de estos grados.

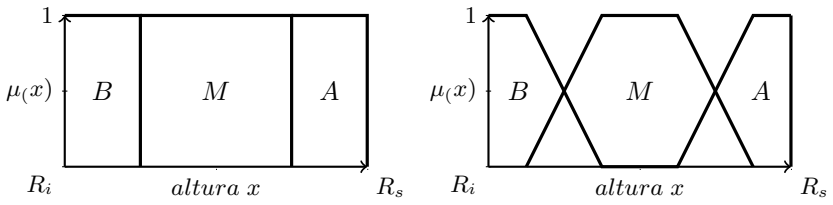


Figura 2.2: Representación del concepto *altura de una persona* mediante una variable categórica-ordinal clásica (Izquierda) y una variable difusa lingüística con función de pertenencia trapezoidal (derecha).

Es posible que esta extensión del solapamiento entre clases dado por la distribución triangular considerada resulte poco realista para la representación del concepto manejado. Así, podría ser adecuado considerar que el grado de solapamiento no se produzca a lo largo del todo el intervalo sino en un rango determinado. Dicho de otra forma, se puede asumir que un individuo presentará un solo grado de pertenencia a una de las clases siempre y cuando su valor numérico se sitúe no solo en el valor (centro o extremo) comentado sino en un entorno más amplio de éste. Por consiguiente, se puede modelar este concepto permitiendo un menor solapamiento (llegando a ser nulo en ciertos intervalos) entre clases adyacentes por medio del uso de funciones de pertenencia de tipo *trapezoidal* como se ilustra en la Figura 2.2.

2.3. Noción de Bipolaridad

Como seres humanos, el conocimiento sobre nuestros propios métodos de razonamiento ha supuesto una fructífera fuente de inspiración en el desarrollo de la Inteligencia Artificial (IA) desde su aparición en la mitad del pasado siglo. De hecho, hoy en día, muchas de las áreas más relevantes de esta disciplina siguen íntimamente relacionadas con la representación de algunos de los procesos y capacidades mentales de los seres humanos. Asimismo, los sistemas expertos y sistemas de ayuda a la toma de decisiones frecuentemente tratan de replicar esa habilidad propiamente humana para la correcta toma de decisiones en muchos campos de aplicación (ciencias de la salud y decisiones médicas, publicidad, tareas de control, gestión de desastres, etc.) basada en un conocimiento explícito sobre dichas realidades concretas. De una forma similar, las técnicas de *Machine Learning* y MDD (ver [44]) han encontrado gran éxito en la replicación de la capacidad de reconocimiento de patrones y aprendizaje basado en experiencia intrínseca al ser humano.

Es entonces necesario destacar que este éxito no habría sido posible sin la búsqueda de inspiración en nuestro propio conocimiento de los procesos para la resolución de problemas en los campos de aplicación concretos y, de ninguna forma se habría avanzado en este aspecto sin el desarrollo de modelos formales apropiados que permitan la representación de este conocimiento.

En este sentido, este trabajo se centra en la búsqueda de inspiración basada en la idea de *bipolaridad* que, como no podía ser de otra forma, es en origen una noción de corte psicológico (ver [11, 61, 86]). Este conceptos está basado concretamente en la capacidad humana para tratar la información disponible para la toma de decisiones por medio de la asignación en dos distintos grupos según su carácter positivo ó negativo para la toma de la decisión final. De esta forma, es habitual el uso de un sistema de razonamiento para la toma de decisiones basado en la consideración de los aspectos positivos y negativos de las alternativas disponibles [20, 28, 47, 62], con una elección final basada en el balance óptimo entre ambos extremos o polaridades. Debido a la existencia de tales extremos o polaridades, esta característica del juicio humano es frecuentemente conocida como *bipolaridad*.

Esta habilidad para lidiar simultáneamente con informaciones de opuesto carácter (positivo y negativo) permite a nuestra mente ser capaz de representar y manejar situaciones complejas en presencia de determinados intereses y emociones como la ambivalencia, la indeterminación o el conflicto. Es por ello que la idea de *bipolaridad* tiene una larga historia. Así, en los últimos 25 años este concepto ha sido un foco de atención de investigadores en el campo de la lógica y la representación de conocimiento. En particular, muchos autores (ver [6, 20, 58, 59]) han enfatizado la relevancia de introducir un enfoque bipolar en el marco de los formalismos propios de la representación del conocimiento como la lógica difusa [87] o la teoría de la posibilidad [21] como medio para mejorar su poder de representación de la realidad.

Se pueden destacar tres nociones básicas en el contexto de la representación del conocimiento, la de *polaridad psicolingüística*, la de *antonimia* y la de *disimilitud*, que han tenido un papel relevante en la formulación de paradigmas en lingüística y en psicología y que han sido objeto de atención por parte de la comunidad científica matemática y de inteligencia artificial en años recientes, conduciendo, por ejemplo, a la introducción del concepto de bipolaridad.

Como se ha mencionado, esta noción de bipolaridad recoge la distinción entre información de carácter positivo y negativo que suele impregnar el raciocinio humano, y que en muchas ocasiones refleja procesos y afectos independientes o al menos no complementarios. La introducción de esa distinción en un modelo de representación del conocimiento conlleva el uso de escalas y estructuras lógicas de mayor complejidad que en el caso no bipolar.

En un contexto lingüístico, es interesante comprobar que los afectos positivos y negativos que condicionan el carácter de casi cualquier información suelen estar asociados, al menos en una primera aproximación, a significados opuestos, y por ende, a vocablos antónimos. Así sucede, por ejemplo, con los pares bueno/malo o verdadero/falso. Sin embargo, en algunos contextos es posible seguir utilizando esta semántica opuesta aun en el caso de no contar con vocablos antónimos, o incluso con ningún vocablo. Esto da pie a

considerar, en lugar de la antonimia [76], la noción más general de antagonismo o disimilitud semántica, propuesta en [65, 66], como base de la idea de bipolaridad.

2.4. Clasificadores nítidos, probabilísticos y difusos

En esta sección se introduce una formalización de distintos modelos de clasificación atendiendo al tipo de salida que proporcionan con el objetivo de motivar el propósito principal de este trabajo: la importancia de modelar la información de naturaleza *soft* obtenida por la gran mayoría de algoritmos de clasificación en la etapa previa a la toma de decisiones.

Así, es posible distinguir entre clasificadores de naturaleza nítida, probabilística y difusa cuyas definiciones formales se presentan a continuación. Sea $S = \{C_1, \ldots, C_c\}$ el conjunto de clases en el problema de clasificación considerado y sea $X = \{x_1, \ldots, x_N\}$ el conjunto de instancias a clasificar.

Como se ha comentado en secciones previas, muchos usuarios de los algoritmos de clasificación tienen en cuenta únicamente la salida final del proceso de clasificación, esto es, la clase final asignada para cada instancia Este hecho se debe probablemente al interés por conocer de manera única la solución final alcanzada por el clasificador. Por esta razón, es pertinente señalar que, de forma general, los clasificadores son entendidos como funciones

$$C : X \longrightarrow \{C_1, \ldots, C_c\}, \tag{2.1}$$

esto es, un procedimiento que asigna una única clase final a cada una de las instancias a clasificar.

Sin embargo, el proceso de clasificación está compuesto de varias etapas antes de tomar la decisión final de asignación a una cierta clase y, es precisamente en estos pasos intermedios donde la información *soft* aparece de manera natural para modelar la evidencia de carácter numérico que representa la fuerza de asociación entre instancias y clases de pertenencia. En particular, es muy común que los algoritmos de clasificación manejen información de carácter *soft* para cada instancia $x \in X$ acerca de la probabilidad de que el elemento x pertenezca a cada una de las distintas clases o, en el contexto de los calificadores de naturaleza difusa, acerca del *grado de pertenencia* de la instancia x a ciertas clases.

Considerando esta realidad, en [80] se distingue entre dos tipos de clasificadores *soft*, los de naturaleza puramente nítida definidos en la Ecuación (2.1) y aquellos con evidencia probabilística. En este sentido, un clasificador probabilístico puede ser interpretado como una función

$$C_P : X \longrightarrow [0, 1]^c, \tag{2.2}$$

que asigna a cada registro o instancia x, su probabilidad de pertenencia

a cada una de las clases disponibles. Obviamente, para cada $x \in X$ ha de satisfacerse que $\sum_{i=1}^{c}(C_P(x))_i = 1$ debido a la propiedad de aditividad inherente al concepto de probabilidad.

Es relevante destacar que muchos algoritmos de clasificación y, en particular los detallados en la Sección 2.5, pueden ser entendidos como clasificadores de naturaleza probabilística atendiendo en exclusiva a la información *soft* proporcionada por dichos modelos en la etapa previa a la asignación nítida final.

Es posible definir de manera análoga un clasificador de naturaleza difusa como una función,

$$C_F : X \longrightarrow [0,1]^c, \tag{2.3}$$

donde ahora la restricción $\sum_{i=1}^{c}(C_F(x))_i = 1$ no es necesariamente satisfecha, es decir, los grados de pertenencia $\mu_{C_i}(x) = (C_F(x))_i$ de un objeto o instancia x a cada una de las clases C_i no han de sumar necesariamente la unidad. Es posible poner de manifiesto que algunas técnicas de clasificación como por ejemplo los Sistemas de Clasificación Basados en Reglas Difusas (SCBRDs) detallados en la Sección 2.6, se pueden entender como clasificadores difusos en el sentido comentado atendiendo a la información *soft* obtenida en la etapa previa a decisión, la cual es identificada en este contexto como un tipo de *defuzzificación*, entendiendo este como el proceso de conversión de la evidencia de tipo difuso en información nítida.

2.5. Algoritmos de clasificación de naturaleza probabilística

En el contexto de la MDD, multitud de esquemas para resolver tareas de clasificación han sido desarrollados. Entre ellos, destacan los clasificadores de naturaleza probabilística definidos en la Ecuación (2.2), en los que las evidencias de asociación entre patrones y clases vienen dadas en forma de probabilidades.

No pretende ser esta sección una de revisión de la literatura existente en el contexto de los algoritmos de clasificación supervisada. No obstante, dado que algunos de entre ellos son utilizados en el desarrollo de las propuestas contenidas en esta memoria, merecen, al menos, unas líneas de dedicación.

Los algoritmos de naturaleza probabilística utilizados en este trabajo, así como una breve descripción de sus características, se dan a continuación:

Árbol de Decisión con metodología CART (Classification And Regression Trees). Los clasificadores basados en árboles son una de las aproximaciones más intuitivas para la resolución de tareas de clasificación. Estos modelos siguen un esquema de partición del espacio de características en regiones que, finalmente, serán asociadas a una de las clases de la variable objetivo. En general, los clasificadores basados en reglas presentan una gran ventaja respecto a otros paradigmas, la facilidad para interpretar los resultados en términos de las características de entrada. De esta forma, el proceso de un árbol de decisión como CART [8] se compone de una búsqueda iterativa de las variables que producen particiones binarias óptimas (en sentido de alguna medida de referencia como el índice de Gini) en cada punto de partición. Así comenzando por el nodo raíz, se van generando particiones dicotómicas recurrentes hasta que se alcance cierto criterio de parada para el crecimiento del árbol. Entonces, el árbol de clasificación, retorna una serie de nodos finales con la menor "impureza", es decir, con la distribución de clases con menor dispersión. Finalmente, se asigna cada item a la clase que resulte la mayoritaria en el nodo final correspondiente. Cabe destacar que cada nodo final se asocia con una regla de clasificación nítida, por ello se puede enmarcar el algoritmo CART, así como todos los clasificadores tipo árbol en la categoría de *Clasificadores Basados en Reglas (CBR)*, que la misma categoría en la que se encuentran los SCBRDs, siendo estos una particularización de aquellos en contexto difuso.

Redes Neuronales (ANN). Las Redes Neuronales [64, 78] aparecieron por primera vez en los años 50 como intento de elaborar una herramienta capaz de resolver de problemas de forma análoga a como lo haría un cerebro animal. Una de las grandes ventajas de las Redes Neuronales es su capacidad de adecuada adaptación a cualquier entrada por compleja que sea mediante un proceso iterativo. El cuerpo de la neurona se representa como un sumador lineal de los estímulos externos y una función no lineal, que se denomina función de activación y es la que utiliza la suma de estímulos para determinar la actividad de salida de la neurona.

A partir de este perceptrón básico se puede llegar a elaborar un perceptrón multicapa combinando varias neuronas conectadas entre sí. En el perceptrón multicapa se consideran las capas de datos input (entradas, estímulos), capa de nodos ocultos (puede ser más de una) y finalmente la capa de salida. Para utilizar una red neuronal con garantías se requieren un número relativamente alto de observaciones. Además, al no haber inferencia propiamente dicha es necesario emplear datos de entrenamiento para validar el modelo. El entrenamiento de una red neuronal, tiene por objetivo modificar iterativamente los pesos sinápticos de la red con el fin de minimizar el error entre la predicción y la respuesta esperada. Dentro de los parámetros que definen una red, la función de red más utilizada es de tipo lineal, y como

función de activación más empleada está la función sigmoidea.

Random Forest (RF) En el contexto de los multiclasificadores, esto es, aquellos cuya decisión final esta basada en el ensamblado de varios clasificadores individuales, el algoritmo Random Forest [29] aprovecha la técnica de *bagging*, consistente en la generación de diversas muestras aleatorias con reemplazamiento de tamaño igual al del conjunto de datos, para incrementar el rendimiento en la clasificación mediante la incorporación de aleatoriedad en la construcción de cada clasificador individual. Esta aleatoriedad en su construcción permite que los clasificadores base estén menos correlacionados, lo que redunda en una reducción de varianza al realizar su combinación. Por este motivo, la metodología de Random Forest proporciona mejores resultados si se emplea sobre clasificadores base con bajo sesgo y varianza relativamente alta, como es el caso de los árboles de clasificación, y en particular de CART, cuando no se imponen límites al crecimiento y no se realiza poda. La componente azarosa del proceso puede introducirse en la partición del espacio (construcción del árbol), así como en la muestra de entrenamiento. El algoritmo Random Forest, a diferencia del bagging selecciona de forma aleatoria en cada ramificación un subcojunto de p variables explicativas entre las disponibles, y de estas selecciona la mejor para realizar la partición. Se presenta a continuación el proceso del algoritmo:

- Selecciona individuos al azar (usando muestreo con reemplazo) para crear diferentes conjuntos de datos sobre los que entrena cada árbol.
- Al crear los arboles se eligen variables al azar en cada nodo del árbol, dejando crecer el árbol en profundidad (sin podar).
- Crea un árbol de decisión con cada set de datos, obteniendo diferentes arboles, ya que cada set contiene diferentes individuos y diferentes variables.
- Predice los nuevos datos usando el *voto mayoritario*, donde la instancia se clasifica en la clase predicha por una mayoría de árboles. Este enfoque puede ser así mismo probabilístico.

Gradient Boosting (GBM) El algoritmo de Gradient Boosting [30] se basa en la idea de entrenar el algoritmo mediante la actualización de los pesos de las observaciones pertenecientes a las clases del suceso de interés a través de la optimización en dirección descendente de una función de pérdida o error determinada, consiguiendo dar mayor relevancia en cada iteración a las observaciones mal clasificadas en pasos anteriores.

2.6. Sistemas de Clasificación Basados en Reglas Difusas (SCBRDs)

Los sistemas difusos constituyen una de las áreas más fructíferas de aplicación de la Teoría de los Conjuntos Difusos. En particular, los SCBRDs pueden ser entendidos como una extensión de los sistemas basados en reglas clásicos ya que se basan igualmente en el uso de predicados de tipo *"SI-ENTONCES"*. Sin embargo, los formalismos de representación de los antecedentes (y en algunos casos, los consecuentes) trascienden la lógica clásica de conjuntos, adquiriendo un carácter más amplio basado en la noción de conjunto difuso (ver Sección 2.2) propuesta por Zadeh en 1965 [87] en un intento de replicar el razonamiento de expertos humanos por medio de reglas capaces de manejar problemas de la vida real en un gran abanico de dominios de aplicación.

En las siguientes secciones se pretende proporcionar una base conceptual preliminar sobre los formalismos propios de los SCBRDs, destacando los distintos tipos de reglas utilizadas y el método de razonamiento difuso.

2.6.1. Conceptos generales

Desde su aparición en la segunda mitad del pasado siglo, son muchas las propuestas para la generación y el aprendizaje de reglas difusas en el contexto de los problemas de clasificación, desde heurísticas simples de búsqueda [1**?** **?**] hasta técnicas neuro-difusas [7, 60] o algoritmos genéticos.

Debido a su capacidad de generación de modelos basados en reglas que manejan adecuadamente términos lingüísticos y que son, por tanto, de fácil interpretación, los SCBRDs son ampliamente utilizados para tareas de MDD [3, 4], incluso es problemas de clasificación de especial dificultad como clasificación con clases desbalanceadas [23] o en entornos de *big data* [24]. Es especialmente relevante hoy en día la discusión acerca de la Inteligencia Artificial Explicable (IAE) [2, 72] y la aplicación de los SCBRDs en este contexto debido a las propiedades comentadas (ver [25, 12, 54, 55]).

Como se enfatizó en la Sección 2.2, los conjuntos y la lógica difusos hacen posible una modelización adecuada del significado de los términos propios del lenguaje natural, necesariamente afectados por una imprecisión semántica. Por ello, las reglas generadas bajo este paradigma son interpretables a un nivel lingüístico.

En el contexto de los SCBRDs, las funciones de pertenencia permiten la evaluación de las variables independientes o explicativas en forma de palabras del lenguaje natural. Las reglas *"SI-ENTONCES"* difusas son, por ende, construidas por medio de la asociación local entre distintas características de las variables explicativas, dadas en términos del lenguaje, y la verificación de alguna de las clases disponibles en el problema de clasificación.

En un marco general, los dos componentes principales de los SCBRDs son los citados a continuación:

- Base de Conocimiento (BC): Compuesta por la Base de Reglas (BR) y la Base de Datos (BD), donde respectivamente se almacenan las reglas y las funciones de pertenencia.
- Método de Razonamiento Difuso (MRD): Es el mecanismo que permite la clasificación de los ejemplos en base a la información contenida en la BC.

Por tanto, es este esquema posibilita una asociación entre la BC de un SCBRD con el modelo que recoge las relaciones entre las variables explicativas y las clases, ya sea por medio de una especificación dada por un experto humano o aprendida de los datos de entrenamiento. Entonces, el MRD especifica las instrucciones sobre cómo utilizar la BC para llevar a cabo el proceso de clasificación de nuevas instancias, esto es, de determinar cómo el proceso de inferencia se realiza a partir del modelo dado.

2.6.2. Tipos de reglas difusas

En un esquema de clasificación clásico, se parte de un conjunto de P patrones etiquetados $x_p = (x_{p1}, ..., x_{pn})$, donde $p \in \{1, 2, ..., P\}$ y x_{pi} representa el valor de la i-ésima característica X_i en el p-ésimo patrón, $i \in \{1, 2, ..., n\}$. Se asume que cada patrón x_p pertenece a una clase C_p, donde C_p es una de las c clases que componen el conjunto de clases $S = \{C_1, ..., C_c\}$ del problema de clasificación supervisada.

Para explotar la información contenida en los datos, un SCBRD ha de convertir las variables numéricas en grados de pertenencia a las distintas etiquetas lingüísticas consideradas. Esta conversión es llevada a cabo por medio de un proceso de *fuzzificación*, en el cual los valores x_{pi} son transformados en vectores $(\mu_{A_i^1}(x_{pi}), \ldots, \mu_{A_i^{l_i}}(x_{pi})) \in [0,1]^{l_i}$, donde l_i es el número de etiquetas lingüísticas consideradas para evaluar la característica i y $\mu_{A_i^k}$ representa la función de pertenencia del k-ésimo término lingüístico empleado para describir la i-ésima variable, $k \in 1, \ldots, l_i$.

Una vez el conjunto de datos ha sido sometido al proceso de *fuzzificación* por medio de la utilización de funciones de pertenencia especificadas en la BD, un SCBRD crea la base de reglas de acuerdo a cierta heurística. En este sentido y como se ha señalado en la Sección 2.6.1, existen diversas formas de crear y aprender reglas en la literatura, desde los más simples procedimientos hasta sistemas difusos evolutivos que hoy en día destacan en el campo de la Inteligencia Artificial Explicable (IAE) [2, 25].

Existe un consenso general sobre la consideración de tres tipos de reglas *"SI-ENTONCES"* difusas [16]:

- Reglas de Tipo I con una clase en el consecuente.

  Regla R^q: SI X_1 is A_1^q y ... y X_n is A_n^q ENTONCES clase C_j, $j \in \{1, 2, ..., c\}$

  donde A_i^q representa el antecedente lingüístico utilizado en la regla q para la i-ésima variable, y C_j en el consecuente denota una de las c clases.

- Reglas de Tipo II con una clase y un grado de certeza r^q en el consecuente.

  Regla R^q: SI X_1 is A_1^q y ... y X_n is A_n^q ENTONCES clase C_j con r^q

  donde $j \in \{1, 2, ..., c\}$ y r^q es el grado de certeza de la regla R^q, usualmente un número real en el intervalo $[0, 1]$.

- Reglas de Tipo III con un grado de certeza para todas las clases en el consecuente.

  Regla R^q: SI X_1 is A_1^q y ... y X_n is A_n^q ENTONCES $(r_1^q, .., r_c^q)$.

Es interesante señalar que las reglas tipo I son un caso particular de las reglas tipo II cuando $r^q = 1$. Análogamente, las reglas tipo II pueden se vistas como una particularización de las reglas tipo III en las que $r_i^q = 0$ si $i \neq j$. De otro modo, una regla de tipo III ser puede entender como la combinación de reglas tipo II con el mismo antecedente pero distintas clases consecuentes. Los grados de certeza o *pesos de las reglas* r_j^q son interpretados habitualmente como una medida de la fuerza de asociación entre la premisa de la regla, dada por la condición $A^q = A_1^q$ *y...y* A_n^q en el espacio de entrada, y la clase C_j.

Muchas son las definiciones heurísticas de estos pesos de las reglas y la mayoría de ellas están basadas en la noción de confianza de una regla difusa, expresada como,

$$cf(A^q \Rightarrow C_j) = \frac{\sum_{p|C^p=C_j} \mu_{A^q}(x_p)}{\sum_p \mu_{A^q}(x_p)}$$

donde $\mu_{A^q}(x_p) = T(\mu_{A_1^q}(x_{p1}), \ldots, \mu_{A_n^q}(x_{pn}))$ y T representan un operador de conjunción como una t-norma o una función de *overlap* n-dimensional [59]. Notar que la confianza de la regla $A^q \Rightarrow C_j$ puede interpretarse como una estimación de la probabilidad condicionada de la clase C_j dada por la condición difusa A^q.

2.6.3. Método de Razonamiento Difuso (MRD)

El segundo componente de un SCBRD es el Método de Razonamiento Difuso (MRD), que determina el proceso de inferencia, responsable de clasificar

nuevas instancias de acuerdo a las relaciones entre cada característica de las instancias y el modelo almacenado en la BC.

Siguiendo a [16], dado un patrón $x = (x_1, ..., x_n)$ a ser clasificado en una de las c clases disponibles en el conjunto de clases S, y asumiendo una BR con N_R reglas, el modelo general del método de razonamiento difuso consta de las siguientes etapas :

- *Grado de emparejamiento* del patrón x y el antecedente de todas las reglas R^q en la BR. Este grado es normalmente obtenido mediante la aplicación de un operador de conjunción T (ya sea una t-norma o una función de *overlap*) a los grados de pertenencia de los valores $x_1, \ldots, x_n$ dentro de los términos lingüísticos $A_1^q, \ldots, A_n^q$ en el antecedente de la regla R^q, por ejemplo,

$$\sigma_{R^q}(x) = T(\mu_{A_1^q}(x_1), ..., \mu_{A_n^q}(x_n)), q = 1, ..., N_R$$

- *Grado de asociación* del patrón x con la clase C_j atendiendo a cada regla R^q, obtenido por agregación (usualmente a través del operador producto), los grados de emparejamiento previamente definidos $\sigma_{R^q}(x)$ del patrón x con el antecedente de la regla y el grado de certeza o peso de la regla r_j^q para la clase C_j en el consecuente de la regla:

$$b_j^q(x) = h(\sigma_{R^q}(x), r_j^q),\ j = 1, .., c,\ q = 1, ..., N_R$$

- *Función de ponderación* en la forma $g : [0, 1] \times [0, 1]$ para favorecer grados de asociación altos y penalizar aquellos mas bajos. Elecciones típicas de para g son funciones de tipo sigmoidal, y la función identidad en caso de no desear ponderaciones.

$$w_j^q(x) = g(b_j^q(x))$$

- *Grado de consistencia del patrón de clasificación* para todas las clases, calculado mediante la aplicación de una función de agregación f que combina, para cada clase C_j, los grados de asociación positivos ponderados $w_j^q(x)$ de todas las reglas calculadas en etapas previas. Quizá las más comunes de entre las posibles elecciones para f sean las funciones máximo y suma, respectivamente dando lugar a los conocidos métodos de *regla ganadora* y *suma ponderada*.

$$\pi_j(x) = f(w_j^q(x), q = 1, ..., N_R),\ j = 1, ..., c$$

- *Clasificación nítida final*, producida a través de un proceso *defuzzificación* que transforma el grado de consistencia de todas las clases en

una única asignación, por medio de la aplicación de una función de decisión F. Con mucho, la más frecuente elección para F es la regla del máximo, que asigna el patrón x a la clase $C_h \in S$ con el grado de consistencia mayor, por ejemplo,

$$C_h = F(\pi_1(x), .., \pi_c(x))$$

tal que

$$\pi_h(x) = \max_{j=1,..,c} \pi_j(x)$$

2.7. Medidas de evaluación en Minería de Datos

En el contexto de la clasificación y la MDD, es de vital importancia la consideración de las medidas que determinan la *calidad* o *rendimiento* de los algoritmos o modelos aplicados a un cierto conjunto de datos. En este sentido y, debido al carácter supervisado de este tipo de aprendizaje, la más intuitiva de entre estas medidas es la tasa de fallos o tasa de aciertos según se mire, la conocida en términos anglosajones como *accuracy*. Esta medida es directamente extraída del cruce de la variable que contiene las clases reales, digamos Y, y aquella que reúne las clases predichas por el modelo, $\hat{Y}$, en lo que se conoce como tabla o matriz de confusión del modelo. En esta matriz de dimensión $c \times c$ se consigna en cada elemento el número de instancias que cumplen simultáneamente la pertenencia a cierta clase C_j en las distribuciones de las variables *real* y *predicha*. De esta forma, una matriz diagonal indica un ajuste perfecto (tasa de acierto o *accuracy* con valor 1) mientras que una matriz con todos los valores no nulos fuera de esta diagonal representa el peor de los escenarios, con una tasa de aciertos nula.

Considérese un problema de clasificación binaria, esto es, con dos clases en la variable objetivo $S = \{C_1, C_2\}$ y N ejemplos a clasificar. Supongamos una variable que identifica la clase real Y y otra que contiene la clase asignada por el clasificador $\hat{Y}$. En esta situación, se puede generar la tabla de contingencia de esta dos variables de naturaleza categórica como sigue,

Tabla 2.1: Distribuciones conjuntas y marginales de las clases reales y predichas $(Y, \hat{Y})$

$Y, \hat{Y}$	C_1	C_2	
C_1	n_{11}	n_{12}	$n_{1.}$
C_2	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	N

En la tabla 2.1, n_{ij}, con $i, j = 1, 2$ representa el número de instancias pertenecientes a la clase C_i que han sido clasificadas como clase C_j. De esta

forma, el clasificador acierta cuando $i = j$, errando en caso contrario. Los valores $n_{i.}$ y $n_{.j}$, representan los totales por filas y columnas, respectivamente. Esto es, $n_{i.} = n_{ii} + n_{ij}$ y $n_{.j} = n_{ij} + n_{jj}$. De estas definiciones se desprenden algunos de los conceptos básicos relativos a la evaluación de tablas de contingencia. En primer lugar se definen las tasas de *verdaderos* (TVP) y *falsos* (TFP) positivos como la proporción de casos de la clase positiva (digamos C_1) que fueron correcta/erróneamente clasificados por el algoritmo con la siguiente expresión:

$$TVP = \frac{n_{11}}{n_{11} + n_{12}}$$

$$TFP = \frac{n_{12}}{n_{11} + n_{12}}$$

Análogamente se pueden definir las tasas de *verdaderos* (TVN) y *falsos* (TFN) negativos como la proporción de casos de la clase negativa (digamos C_2) correcta/incorrectamente clasificados por el modelo mediante la siguiente ecuación:

$$TVN = \frac{n_{22}}{n_{21} + n_{22}}$$

$$TFN = \frac{n_{21}}{n_{21} + n_{22}}$$

De los conceptos anteriores puede extraerse la definición de la tasa de aciertos o *accuracy* (Acc) como,

$$Acc = \frac{Verdaderos\ Positivos\ +\ Verdaderos\ Negativos}{N} = \frac{n_{11} + n_{22}}{N}$$

o de una forma general,

$$Acc = \frac{\sum_{i=1}^{c} n_{ii}}{\sum_{i=1}^{c} \sum_{j=1}^{c} n_{ij}} = \frac{\sum_{i=1}^{c} n_{ii}}{N} \tag{2.4}$$

Esta medida, aunque razonable, no está ni de lejos exenta de inconvenientes cuya existencia ha de provocar cierta alerta en la consideración de este valor numérico como criterio único de evaluación de modelos o algoritmos de clasificación. Así, en el caso de conjuntos de datos altamente desbalanceados, esto es, cuando una clase está poco representada o presenta una prevalencia baja, la decisión basada en este criterio puede resultar peligrosa debido a la naturaleza misma de la medida.

Ejemplo ilustrativo . Se puede imaginar un problema de clasificación binaria con $C = (C_1, C_2)$ y $N = 100$ ejemplos a clasificar, en el que ahora $n_2. \ll n_1.$. En esta situación, se puede considerar la matriz de confusión dada en la Tabla 2.2.

Tabla 2.2: Distribuciones conjuntas y marginales de las clases reales y predichas $(Y, \hat{Y})$ en clasificación de conjuntos de datos desbalanceados.

$Y, \hat{Y}$	C_1	C_2	
C_1	95	0	**95**
C_2	5	0	**5**
	100	**0**	**100**

En este ejemplo y atendiendo únicamente al valor de *accuracy*, determinaríamos que el clasificador en cuestión, con $Acc = 0.95$ obtiene un resultado de clasificación considerado como excelente. Sin embargo, como se observa en la Tabla 2.2, no se produce acierto alguno en la clasificación de instancias de la categoría poco representada, la cual suele ser la clase de interés del problema de clasificación (como en el caso del problema de estudio de siniestralidad vial en el que la tasa de representación de la clase de mayor interés, *resultado mortal*, es inferior al 3 %). Esto hace necesaria la consideración de otras medidas como criterio de evaluación y comparación de modelos en el contexto de la clasificación supervisada que tengan en cuenta estos aspectos.

En este punto, la elección del estadístico *kappa* de Cohen parece una de las mas razonables debido a que esta métrica compara el *acuerdo observado* con el *acuerdo esperado* bajo clasificación al azar, evitando así el inconveniente comentado.
El estadístico kappa se define formalmente como,

$$\kappa = \frac{p_0 - p_c}{1 - p_c} \tag{2.5}$$

donde p_0 representa el acuerdo observado y p_c el acuerdo esperado bajo asignación aleatoria. En el ejemplo binario de la Tabla 2.1 se tiene que,

$$p_0 = \frac{\sum_{i=1}^{c} n_{ii}}{N} \quad p_c = \frac{\sum_{i=1}^{c} n_{i.} n_{.i}}{N}$$

Esta medida es utilizada no solo para evaluar la acuerdo de clasificadores únicos sino también para la comparación de distintos algoritmos Adicionalmente, tiene en cuenta la clasificación al azar, lo que generalmente resulta menos engañoso que la utilización de la tasa de aciertos o *accuracy* como medida. Así, un *acuerdo observado* del 80 % es mucho menos impactante con un *acuerdo esperado* del 75 % que con un *acuerdo esperado* del 50 %. El cómputo de *acuerdo observado* y *acuerdo esperado* es esencial para la comprensión del estadístico kappa.

Los clasificadores construidos y evaluados en conjuntos de datos con distintas distribuciones de clases pueden ser comparados con mayor fidelidad a través de la métrica kappa que por medio de la tasa de aciertos o *accuracy* debido a su escalamiento en relación a la acuerdo esperada. Por tanto, kappa resulta un mejor indicador del grado de acierto del algoritmo a través de todas las instancias y clases, ya que el *accuracy* podría estar sesgado si la distribución de clases lo estuviera.

Considerando de nuevo el ejemplo mostrado anteriormente, en el que la utilización de la métrica *accuracy* se demostró engañosa, el estadístico kappa es capaz de reflejar con mayor fidelidad el comportamiento del clasificador en el contexto de desbalanceo Se tiene que $p_0 = Acc = 0.95$ y de la misma forma resulta que $p_c = 0.95$, es decir, el acuerdo al azar y el dado por el clasificador coinciden o, dicho de otro modo, el algoritmo considerado no produce mejora alguna con respecto a la clasificación aleatoria y por tanto, como es deseable, la métrica refleja un valor de $\kappa = 0$ indicando la baja calidad del modelo.

Cabe destacar que en el caso de clases balanceadas las métricas kappa y *accuracy* coinciden en valor por lo que no existe ninguna desventaja en el uso del estadístico kappa como criterio de evaluación de algoritmos en el contexto de la clasificación supervisada.

2.8. Metaheurísticas de búsqueda

Esta sección describe brevemente las metaheurísticas de búsqueda aplicadas en este trabajo para encontrar la estructura de disimilitud óptima entre el conjunto de clases. En efecto, la utilización de metaheurísticas de búsqueda para la selección de parámetros en el contexto de la clasificación supervisada ha adquirido un carácter generalizado en los últimos años [4, 7, 23, 25, 44, 73], tanto aquellas técnicas genéticas inspiradas en el raciocinio humano (Algoritmos Genéticos ó AG [41, 52, 53]) como las basadas en los patrones de comportamiento observados en la naturaleza (Algoritmos Inspirados en la Naturaleza).

Algoritmos Genéticos (AG) Los algoritmos genéticos (AG) [39, 41] son un modelo de la evolución biológica basada en la teoría de la selección natural de Charles Darwin. Fue probablemente el primer modelo en utilizar el cruce y la recombinación, mutación y selección en el estudio de sistemas adaptativos y artificiales. Estos operadores genéticos forman la parte esencial del algoritmo genético, así como una estrategia de resolución de problemas. Desde entonces, se han desarrollado muchas variantes de algoritmos genéticos y se han aplicado a una amplia gama de problemas de optimización.

Hay muchas ventajas de los algoritmos genéticos sobre los algoritmos de optimización tradicionales. Dos de los más notables son la capacidad de lidiar con problemas complejos y en entornos paralelos. Los algoritmos genéticos

pueden tratar varios tipos de optimización, ya sea que la función objetivo o *fitness* sea estacionaria o no estacionaria (cambia con el tiempo), lineal o no lineal, continua o discontinua, o con ruido aleatorio. Debido a que múltiples descendientes en una población actúan como agentes independientes, la población (o cualquier subgrupo) puede explorar el espacio de búsqueda en muchas direcciones simultáneamente. Esta característica lo hace ideal para paralelizar los algoritmos de implementación. Se pueden manipular distintos parámetros e incluso distintos grupos de cadenas codificadas al mismo tiempo.

Sin embargo, los algoritmos genéticos también tienen algunas desventajas. La formulación de una función de *fitness*, el uso del tamaño de la población, la elección de parámetros importantes como la tasa de mutación y el cruce, y los criterios de selección de la nueva población deben realizarse con cuidado. Cualquier elección inapropiada dificultará la convergencia del algoritmo o simplemente producirá pobres resultados. A pesar de estos inconvenientes, los algoritmos genéticos siguen siendo uno de los algoritmos de optimización más utilizados en la optimización no lineal moderna.

Optimizador de Lobos Grises (GWO) GWO es un algoritmo de optimización relativamente nuevo propuesto en [57]. Este algoritmo bio-inspirado imita el mecanismo de caza y la jerarquía de liderazgo de los lobos grises.

En particular, emplea cuatro tipos de lobos grises, a saber, alfa, beta, delta y omega para simular la jerarquía de liderazgo. En orden descendente, los lobos alfa son responsables de tomar decisiones sobre la caza, el lugar para dormir y las actividades relacionadas con el grupo. En segundo lugar, los lobos beta ayudan a los alfas en sus tareas y tienen que reforzar las órdenes de los alfa en la manada y dar retroalimentación a los alfas. Son los mejores candidatos para convertirse en alfas en caso de muerte o retiro de un lobo alfa. El último eslabón de la cadena son los lobos omega, que juegan el papel de chivo expiatorio y son los últimos autorizados a comer. El resto de lobos se llaman subordinados o deltas. GWO es similar a otras metaheurísticas donde la búsqueda comienza con una población de lobos generada aleatoriamente.

Respecto al mecanismo de caza, está compuesto por los siguientes fases: búsqueda, acorralamiento y ataque de la presa. La técnica de caza y la jerarquía social de los lobos grises se modelan matemáticamente para diseñar el GWO y llevar a cabo la optimización.

Capítulo 3

Representación bipolar del conocimiento en el contexto de clasificación supervisada

Jamás dejes que las dudas paralicen tus acciones. Toma siempre todas las decisiones que necesites, incluso sin tener la seguridad o certeza de estar decidiendo correctamente.

Paulo Coelho

Resumen: Este capítulo esta reservado a la presentación de los conceptos relativos a la bipolaridad y la representación del conocimiento en el contexto de la clasificación supervisada desde la perspectiva de planteamiento de este trabajo. En primer lugar, en la Sección 3.1 se discute el concepto de disimilitud y su consideración sobre el conjunto de clases de un problema de clasificación. Tras ello, en la Sección 3.2 se plantea el concepto de *evidencia negativa* y la propuesta de generación de dicha evidencia considerada en esta memoria. La Sección 3.3 contiene las propuestas de explotación de la información de carácter bipolar de cara a la clasificación final, distinguiendo entre simples agregaciones o métodos de explotación más flexibles como los basados en reglas. En la sección 3.4 se presenta un análisis de ciertos tipo de estructuras de disimilitud con el objetivo de comprender mejor su comportamiento en este contexto. Finalmente, como una de las principales aportaciones de este trabajo, se presenta en la Sección 3.5 el método de aprendizaje de la estructura de disimilitud basado en los datos.

3.1. Concepto de disimilitud

En términos generales, en el campo de la clasificación de datos, se asume que las clases son de alguna forma entidades independientes y, por tanto, carecen de relación. Esta asunción puede no reflejar la realidad en muchos campos de aplicación debido en esencia al carácter natural del conjunto de clases. De esta forma, en el campo de aplicación aquí estudiado (el que estudia la siniestralidad vial y las consecuencias de los accidentes de tráfico) así como en aquel estudiado en [68] (sobre las consecuencias de desastres naturales), el conjunto de clases a clasificar no puede entenderse como un conjunto de entidades independientes en tanto existe, de forma orgánica, una relación de similitud única entre pares de clases. En otras palabras, al tratar con conjuntos susceptibles de ordenación, la similitud entre clases adyacentes ha de ser superior que entre clases separadas. Para ilustrar este hecho, se puede pensar en la similitud entre los pares de clases *Ileso* (I) - *Muerto* (M) y *Herido Grave* (HG) - *Muerto* (M). En esta situación, es evidente que los accidentes con consecuencias mortales y graves han de presentar una similitud mayor que la existente entre entre el primer par (ileso-muerto). Es por ello deseable, tener en cuenta esta estructura inherente a la realidad de estudio de cara a una mejor representación del conocimiento contenido en los datos.

La existencia de una estructura de relaciones en el conjunto de clases, no se limita a la consideración de variables susceptibles de ordenación comentadas anteriormente, sino también a aquellas de naturaleza puramente nominal afectadas por una serie de relaciones inherentes a su estructura. Este es el caso, por ejemplo, conjuntos de clases que modelan grupos de individuos en la sociedad. Es difícil asumir que, de manera general, los grupos sociales son entidades independientes per se, exentas de una estructura de relaciones. Así, es claro que determinados pares de grupos de individuos en la sociedad presentan características compartidas o cercanas mientras que otros pares se sitúan en posiciones más alejadas en cuanto a su similitud o características compartidas. Es por ello, que los métodos de modelización de variables de naturaleza ordinal, resultan insuficientes para la representación de este tipo de realidades, haciendo necesaria la consideración de una estructura de clases de carácter más amplio.

En este sentido, muchos son los autores de distintas áreas como teoría de conjuntos difusos [58], Minería de Datos (MDD) [31], operadores de agregación [5, 37, 71] o representación del conocimiento [59], que han resaltado la importancia de tomar en consideración y hacer explícita la estructura de clases inherente que subyace en muchos problemas debido a que esta modelización permite una mejor comprensión de la realidad bajo estudio y un reflejo más fidedigno de la adecuación de los modelos utilizados para la representación de esas realidades. Como se ha venido señalando a lo largo del discurso de este trabajo, en clasificación supervisada resulta usual asu-

mir que las clases son de alguna forma entidades independientes y no son muchos los modelos basados en la consideración de cierta estructura en el conjunto de clases. Sin embargo, la aplicación de clasificadores en diversos contextos, puede verse beneficiada por la introducción de cierta estructura de relaciones entre las clases, de forma que esta relación modelada mejore la adaptación de los clasificadores a las características y requerimientos del propio contexto de aplicación y, por ende, su precisión.

Esta idea fue satisfactoriamente aplicada en [68, 70], donde el conjunto de clases que representa distintos evaluaciones lingüísticas sobre las consecuencias de desastres naturales, fue dotado de una estructura (predefinida) capaz de reflejar la disimilitud natural y las relaciones de oposición de tales evaluaciones. Como consecuencia, se propuso un modelo de clasificación bipolar difuso, en el cual, la asociación entre elementos (escenarios de desastres en este caso) y clases es evaluada a través de pares de grados de pertenencia positivos y negativos, haciendo posible la introducción de algunos requerimientos relativos a la decisión en el contexto de la gestión de desastres.

En este trabajo, como se detallará en la Sección 3.5, uno de los conceptos aportados de mayor relevancia es la extensión de este enfoque bipolar en un marco más amplio, en el cual se asume la existencia de una estructura de disimilitud en el conjunto de clases pero, a diferencia de la aproximación anteriormente comentada en la que se predefine la estructura disímil, ahora se permite que ésta sea aprendida de los datos en el proceso de entrenamiento del modelo. La idea es, por tanto, descubrir la estructura de disimilitud más conveniente en términos de precisión del modelo final.

La extensión propuesta considera como punto de partida un clasificador de tipo *soft* (ya sea probabilístico, difuso o de otra naturaleza) C_S que puede ser dotado de un carácter bipolar por medio de la aplicación del proceso descrito en este capítulo, con grados de evidencia positivos y negativos. Para este propósito, se interpreta el vector de información o evidencia *soft* $ev(x) = (ev_{C_1}(x), \ldots, ev_{C_c}(x))$ dado por el clasificador en forma de puntuaciones numéricas que miden la asociación entre cada objeto y cada clase, como evidencia positiva sobre la asociación entre cada instancia x y cada clase $C_j, j \in \{1, \ldots, c\}$, por lo que se denota este vector en la forma $ev^+(x) = \left(ev^+_{C_1}(x), \ldots, ev^+_{C_c}(x)\right) = (ev_{C_1}(x), \ldots, ev_{C_c}(x))$.

En este punto, se asume que aquellas clases que son disímiles a una cierta clase C_i, en conjunto definen una clase abstracta que puede denotarse por dC_i, que puede ser entendida como "disímil a C_i". Por tanto, sea $ev^-_i(x)$ el grado de evidencia hacia la clase dC_i. Este grado es, obviamente interpretado como evidencia negativa del modelo, en contraposición a la asociación entre la instancia x y la clase C_i ya que proporciona la evidencia sobre la asociación de x con clases disímiles a C_i.

Para la construcción de este grado de evidencia negativa es necesario contar con información adicional que determine cuáles de las clases y hasta

qué grado, son disímiles a cada clase C_i. Se asume que esta relación de disimilitud valorada es irreflexiva (ninguna clase es disímil a sí misma) y no necesariamente simétrica (una clase C_j puede ser disímil a otra C_i sin que ésta última lo sea hacia la primera). Denotamos esta relación por $D = (d_{ij})$, donde $d_{ij} \in [0, 1]$ representa el grado hasta el cual cada clase C_j es disímil a C_i. Esto supone que, cualquier matriz de disimilitud D es, en general no simétrica y debe verificar $d_{ii} = 0$ para todo $i = 1, \ldots, c$. Debido a que informa acerca de las relaciones de disimilitud entre las clases disponibles, se considera que tal matriz D dota al conjunto de clases de una estructura de disimilitud.

Una vez definida la estructura de disimilitud, lo que resta establecer es la formas de generación de evidencia negativa (Sección 3.2) y explotación de los pares de evidencias positivas y negativas (Sección 3.3). Establecidas las convenciones utilizadas, se analizan en la Sección 3.4 estas estructuras de disimilitud y, en particular, su capacidad de ajuste en relación al valor de la evidencia *soft* de base dada por el clasificador, en el marco de las dos agregaciones definidas en 3.3.1, con el objetivo de arrojar cierta luz sobre el comportamiento de distintas estructuras de disimilitud en su aplicación en términos del esquema propuesto en este trabajo.

3.2. Generación de la evidencia negativa

Una forma natural de obtener los grados de evidencia negativa $ev_i^-(x)$ es considerar que proceden de la aplicación de una cierta estructura de disimilitud a los grados de evidencia positiva $ev_j^+(x)$, $j \neq i$. A pesar de que estos grados negativos pueden ser construidos de distintas formas, en este trabajo se considera la siguiente definición, donde D_i denota la i-ésima fila de la matriz disímil D considerada:

$$ev_i^-(x) = \sum_{j \neq i} d_{ij}\ ev_j^+(x) = \sum_{j=1}^{c} d_{ij}\ ev_j^+(x) = D_i ev^+(x) \tag{3.1}$$

o en forma matricial,

$$ev^-(x) = D ev^+(x) \tag{3.2}$$

Toda vez los grados positivos y negativos han sido obtenidos para cada instancia y cada clase mediante los vectores $ev^+(x)$ y $ev^-(x)$, es posible definir un esquema de clasificación bipolar como

$$C_{bip} : X \longrightarrow [0, 1]^c \times [0, 1]^c, \tag{3.3}$$

cuantificando las evidencias positiva y negativa o, de otra forma, a favor o en contra de la asociación entre cada posible instancia $x \in X$ y cada una de las clases disponibles.

Es habitual en el contexto de la clasificación supervisada la asignación de cada ejemplo a una sola clase. En este marco, si una asignación de naturaleza nítida ha de ser propuesta, la información contenida en los vectores $ev^+(x)$ y $ev^-(x)$ deberá ser explotada para decidir la clase a asignar.

3.3. Métodos de explotación de la evidencia bipolar

En el proceso de construcción de sistemas de clasificación como los presentados en esta memoria, existen un gran número de decisiones a tomar en cuanto a la naturaleza, tipo y valor de parámetros, funciones y operadores a utilizar entre otros. Concretamente, es de especial relevancia en el proceso bipolar propuesto la elección de un método para generar la clasificación final de tipo *nítido* en base a las evidencias descritas anteriormente.

Indudablemente, se pueden encontrar en la literatura gran cantidad de posibilidades para enfrentar esta tarea. Especial interés tienen los operadores de agregación [37, 56] y las modernas funciones de *overlap* [10**?** , 38]. Otros métodos de asignación pueden ser, asimismo, considerados atendiendo a la naturaleza del problema a resolver. Es fácil pensar en esta tarea desde un enfoque basado, por ejemplo, en reglas de decisión creadas por un sistema de búsqueda de patrones de relación entre objetos y clases.

Se detallan en esta sección los esquemas de explotación de información bipolar, dada por pares de evidencias positiva y negativa, utilizados en este trabajo con el objetivo de realizar la asignación final a la clases. Desde un enfoque agregativo, se consideran algunos de los operadores diseñados a este efecto como los introducidos en la Sección 3.3.1. Por otro lado se explora en la Sección 3.3.2 una nueva vía de explotación de los pares de evidencias que está basado en reglas de asociación a través de un clasificador tipo árbol como CART.

3.3.1. Operadores de agregación

Abordemos en esta sección la cuestión de cómo agregar, para una clase dada C_i y una instancia x, el par de grados de evidencia positiva y negativa $ev_i^+(x)$ y $ev_i^-(x)$ con el objetivo de obtener un único grado de evidencia ajustado $ev_i^{adj}(x)$. Es obvio que distintos tipos de agregaciones proporcionan grados ajustados diferentes y, por ende, clasificaciones dispares. En este trabajo se han estudiado los dos tipos de agregación definidos a continuación, considerando $ev_i^{adj}(x) = ev_i^{add}(x)$ y $ev_i^{adj}(x) = ev_i^{log}(x)$,respectivamente.

Definición 3.3.1. Sean $ev_i^+(x)$, $ev_i^-(x)$ los grados de evidencia positiva y negativa de la instancia x hacia la clase C_i. El grado ajustado aditivo de x hacia la clase C_i se define como

$$ev_i^{add}(x) = \text{máx}\{0, ev_i^+(x) - ev_i^-(x)\}. \tag{3.4}$$

Cabe destacar que esta definición puede ser interpretada como una t-norma de Lukasiewicz $W(a,b) = \text{máx}\{a+b-1, 0\}$ de los grados positivo y no negativo, esto es, $ev_i^{add}(x) = W(ev_i^+(x), n(ev_i^-(x)))$, donde n es la negación estándar $n(a) = 1-a$. Por tanto, la agregación aditiva $ev_i^{add}(x)$ proporciona el grado en que la asociación de un elemento y una clase reúne evidencia positiva y no negativa, llevando a cabo este proceso mediante el balance de los grados bipolares bajo un esquema aditivo.

De esta forma, la evidencia positiva $ev_i^+(x)$ inicialmente proporcionada por el clasificador es ajustada mediante la sustracción de la evidencia negativa $ev_i^-(x)$. Como caso particular, los grados de evidencia iniciales no se ven modificados cuando ninguna clase es disímil a la clase C_i, o de otra forma, cuando $D_i = 0$.

Por tanto, un grado ajustado aditivo $ev_i^{add}(x) > 0$ representa la existencia de un desajuste entre el soporte para la case C_i y el soporte para la clase dC_i, es decir, para las clases consideradas como disímiles a C_i. En tal situación, la fuerza de la asociación de la instancia x con la clase C_i puede verse reducida en relación a su evaluación inicial, siendo aún perfectamente posible la asignación a la misma clase C_i. Contrariamente, un valor nulo de evidencia ajustada aditiva $ev_i^{add}(x)$ indica una situación en la que existe un mayor grado de evidencia para la clase disímil dC_i que para la propia clase C_i, por lo que el clasificador no debe asignar el ítem a dicha clase.

En la siguiente definición se propone un método alternativo para la agregación de las evidencias positiva y negativa en un único grado ajustado.

Definición 3.3.2. Sean $ev_i^+(x)$, $ev_i^-(x)$ los grados de evidencia positiva y negativa de la instancia x hacia la clase C_i. El grado de evidencia ajustado logístico de x hacia la clase C_i se define como

$$ev_i^{log}(x) = \begin{cases} 1 - e^{-\frac{ev_i^+(x)}{ev_i^-(x)}} & si\ ev_i^-(x) > 0 \\ 1 & en\ otro\ caso \end{cases} \tag{3.5}$$

De manera opuesta a la agregación aditiva anteriormente definida, esta agregación logística está basada en el ratio entre informaciones positiva y negativa, ajustando éste al rango $[0,1]$ a través de una transformación logística. Este esquema permite un comportamiento de alguna forma más flexible de los grados ajustados. En este sentido, la elección de la matriz D puede tener una influencia incluso mayor en el ajuste de los grados de evidencia positiva proporcionados por el clasificador de base considerado, hasta tal punto que puede resultar $ev_i^{log}(x) = 1$ cuando no se reúne evidencia alguna hacia la clase disímil dC_i. En otras palabras, cuando $ev_i^-(x) = 0$. Es por esta razón que la agregación logística aquí propuesta no resulta, a priori, adecuada para trabajar en presencia de estructuras de disimilitud extremas en las cuales

alguno de los d_{ij} es nulo. Sin embargo, cuando no es este el caso (y el algoritmo de aprendizaje de la estructura disímil asegurará rápidamente que no lo es), el aumento de la flexibilidad proporcionada por el operador logístico puede permitir que la clasificación ajustada sea capaz de refinar, de manera efectiva, la famosa regla del máximo inicialmente considerada. Se extiende de esta forma la toma de decisiones para manejar un espectro más amplio de situaciones.

Finalmente, el clasificador bipolar se ve obligado a tomar una decisión nítida acerca de la pertenencia de cada instancia a cada clase debido a la falta de un espectro más amplio de metodologías y medidas de comparación de la precisión de los clasificadores. En este sentido, se considera relevante la creación de un marco de referencia de comparación en el contexto de la clasificación en MDD, en términos que transciendan la comparación nítida basada en medidas relativas a la matriz de confusión de la clasificación.

En cualquier caso, una vez que se ha aplicado uno de estos dos métodos de agregación y obteniéndose los grados ajustados $ev_i^{adj}(x)$ para cada clase (ya sea $ev_i^{adj}(x) = ev_i^{add}(x)$ o $ev_i^{adj}(x) = ev_i^{log}(x)$), la decisión final sobre la clasificación del objeto x se hace aplicando la regla del máximo a dichos grados ajustados. Por lo tanto, el elemento x finalmente se asigna a la clase C_h con un grado máximo ajustado $ev_h^{adj}(x)$, es decir, $h = arg\,\text{máx}_{i\in\{1,\ldots,c\}}\, ev_i^{adj}(x)$.

Para mostrar cómo se modela una salida de clasificación difusa de manera bipolar y mediante una matriz de disimilitud particular, y cómo estas puntuaciones bipolares se agregan en una puntuación difusa ajustada, presentamos el siguiente ejemplo.

Ejemplo 1. Dado un problema de clasificación con variable objetivo binaria $S = (C_1, C_2)$, supongamos que tenemos una instancia x con el siguiente vector de grados de evidencia $ev(x) = (0.6, 0.4)$ tras el proceso de entrenamiento del clasificador considerado. En un esquema de clasificación estándar, mediante el uso de la regla del máximo, esta instancia sería asignada a la clase C_1.

Una vez aplicadas las ideas descritas en esta sección, es posible construir la evidencia positiva y negativa a través de una matriz de disimilitud. Consideremos una situación no simétrica en la que la clase C_1 resulta ser altamente disímil respecto a la clase C_2, no siendo necesariamente cierto el recíproco. Este hecho puede darse en muchas situaciones reales como clasificación de conjuntos desbalanceados o algoritmos de clasificación con errores no simétricos, entre otros.

Si definimos la matriz de disimilitud D con $d_{12} = 1$ y $d_{21} = 0.1$, es posible calcular la evidencia negativa como sigue,

$$ev^-(x) = Dev^+(x) = (d_{12}ev_2^+(x), d_{21}ev_1^+(x)) = (0.4, 0.06)$$

Se observa que la evidencia negativa de esta instancia x es mayor para la

clase C_1 que para C_2. Esto se debe a que la clase C_2 es fuertemente disímil a la clase C_1, así como al hecho de que la evidencia original de la clase C_2 (0.4) no es mucho más baja que de la clase C_1 (0.6). Entonces, los pares de evidencia bipolar para ambas clases son

$$ev_1^{bip}(x) = (ev_1^+(x), ev_1^-(x)) = (0.6, 0.4)$$
$$ev_2^{bip}(x) = (ev_2^+(x), ev_2^-(x)) = (0.4, 0.06)$$

Una vez obtenidos los pares de evidencias positiva y negativa es posible aplicar los dos operadores de agregación anteriormente para obtener los grados de evidencia ajustados siguientes

$$ev^{add}(x) = (0.2, 0.34)$$
$$ev^{log}(x) = (1 - \exp(-\frac{0.6}{0.4}), 1 - \exp(-\frac{0.4}{0.06})) = (0.77, 0.99)$$

dando lugar a la respectiva asignación de clases

$$C_h^{add}(x) = C_h^{log}(x) = C_2.$$

Por consiguiente la estructura de disimilitud considerada es capaz de cambiar la clasificación final dada a la C_1 hacia la clase C_2 por medio de la aplicación de ambas agregaciones.

3.3.2. Funciones de explotación

Pese al buen comportamiento de los esquemas agregativos estudiados hasta ahora, existe una importante limitación en su utilización bajo ciertas realidades. De este modo, el proceso de aprendizaje que se propone en este trabajo (ver Sección 3.5) en conjunción con el uso de las agregaciones anteriores para obtener un grado ajustado que será tratado de forma clásica con una asignación final de clase basada en la regla del máximo (en este punto esta asignación final resulta irrelevante para explicar esta limitación), encuentra una seria restricción en algunas situaciones.

Tal es el caso de la aplicación sobre algoritmos que padecen el conocido y tan temido sobreajuste, que sucede cuando un clasificador aprende con demasiada profundidad las relaciones que se dan en el conjunto de entrenamiento (*train*). Esta afección se traduce en una pobre capacidad de generalización de los patrones encontrados en los datos a nuevas instancias o ejemplos contenidos en el conjunto de prueba (*test*). Es posible identificar estos casos basándonos en las diferencias de rendimiento de los clasificadores en los conjuntos de entrenamiento y prueba.

Por tanto, se ilustra este hecho atendiendo a un ejemplo de clasificación dado por un algoritmo *soft* una vez ha sido entrenado y ajustado a los datos de entrenamiento. En un caso de sobreajuste, se pueden presentar unos valores de rendimiento (consideremos el *accuracy* clásico) como $Acc_{train} = 1$ y $Acc_{test} = 0.65$. Así, la clasificación obtenida en el conjunto de entrenamiento es perfecta por lo que el clasificador ha aprendido los patrones exactos contenidos en este conjunto. No obstante, la capacidad de generalización de este modelo sobre el conjunto de prueba es del todo mejorable. Sería deseable, por tanto, la consideración de un sistema capaz de realizar ajustes adecuados bajo este marco. En el paradigma agregativo, el aprendizaje se ve limitado pues se realiza en base a una función objetivo que es, precisamente, la capacidad de ajuste dada en entrenamiento (por ejemplo el considerado *accuracy*), por lo que el margen de mejora es inexistente.

Como método para soslayar esta limitación, se propone un marco de explotación de la información dada por pares de evidencias positiva y negativa con mayor flexibilidad para enfrentar estos casos extremos. En este sentido, la consideración de un esquema de explotación que trascienda la agregación parece adecuada. De este modo se puede considerar una *función general de explotación* Φ susceptible de aplicación sobre los pares de evidencia bipolar como se muestra en la Definición (3.3.3).

Definición 3.3.3. Sean $ev^+ = (ev_1^+, ..., ev_c^+)$ y $ev^- = (ev_1^-, ..., ev_c^-)$ las evidencias positiva y negativa proporcionada por el clasificador base y Φ una función general de explotación. Entonces la clase asignada a una instancia x, viene determinada por la predicción alcanzada por *función general de explotación* Φ, aplicada sobre los vectores de evidencias positivas y negativas dados por el clasificador base para esa misma observación $ev^+(x) = (ev_1^+(x), ..., ev_c^+(x))$ y $ev^-(x) = (ev_1^-(x), ..., ev_c^-(x))$.

$$Clase(x) = C_i \;\; si\ y\ solo\ si \;\; \Phi(ev^+(x), ev^-(x)) = C_i,\ i \in \{1, \ldots, c\} \quad (3.6)$$

Esto es, la asignación de clase se determina mediante la aplicación de la *función general de explotación*, Φ, sobre las evidencias positivas y negativas.

Por tanto, sería posible considerar una particularización de este marco de explotación general, teniendo en cuenta un sistema basado en reglas que será el encargado de tomar la decisión final de asignación a clases en base a la información contenida en los pares. Se puede pensar en este proceso como la aplicación de un clasificador de naturaleza puramente nítida (ver Ecuación 2.1) a un problema de asignación en el que las variables están compuestas por las evidencias positivas y negativas para cada clase. Se detallan los pormenores de esta propuesta en el Capítulo 7.

3.4. Sobre el comportamiento de las estructuras de disimilitud

Es evidente que la matriz de disimilitud juega un papel fundamental en el proceso de clasificación propuesto ya determina la froma en que la evidencia negativa es generada. Por consiguiente, el rendimiento y los resultados de incorporar información de carácter bipolar al clasificador tienen una dependencia total de la elección de dicha matriz.

Con el objetivo fundamental de analizar el comportamiento de las matrices de disimilitud, en primer lugar se estudia la relación entre la elección de matrices de disimilitud y la decisión final, comenzando por el caso binario con $c = 2$ en un contexto de clasificadores probabilísticos (ver Sección 2.5).

3.4.1. Comportamiento bajo agregación aditiva

Supongamos que se dispone de un sistema de clasificación con dos clases $S = \{C_1, C_2\}$. Para un clasificador *soft* C_S y una instancia x dados, se tiene el vector de evidencias positivas (en este caso probabilidades) asociadas a cada una de las dos clases $C_P(x) = (p_1^+, p_2^+)$ que pueden ser representadas sin pérdida de generalidad como $C_P(x) = (p, 1-p)$. Así mismo, se asume que $p \in (0, 1)$ y, como consecuencia $(1-p)$ estará restringido al mismo intervalo. El caso en que $p \in \{0, 1\}$ puede ser analizado de forma trivial.

La cuestión que merece ser explorada es, por tanto, ¿en qué circunstancias la agregación aditiva es capaz de cambiar la asignación final de clase llevada a cabo por un determinado clasificador? Esta cuestión puede formularse como sigue. Si $p > 0.5$ (y, por ende, una clasificación estándar clásica asignaría la instancia a la clase C_1) ¿cómo debería ser la matriz disímil para mover la instancia a la clase C_2?

Proposición. En un sistema de clasificación binaria con dos clases $S = \{C_1, C_2\}$ y un clasificador probabilístico C_P, sea x un objeto con $C_P(x) = (p, 1-p)$. Entonces se satisface lo siguiente:

1. Si $p < \frac{1}{3}$, entonces $C^{add}(x) = C_2$ para cualquier matriz de disimilitud D.

2. Si $p > \frac{2}{3}$, entonces $C^{add}(x) = C_1$ para cualquier matriz de disimilitud D.

3. Si $p \in [\frac{1}{3}, \frac{2}{3}]$, entonces para todo $j \in \{1, 2\}$ existe una matriz de disimilitud D con la cual $C^{add}(x) = C_j$.

Demostración. Dada cierta matriz de disimilitud $D = \begin{pmatrix} 0 & d_{12} \\ d_{21} & 0 \end{pmatrix}$. El vector p^- de probabilidades negativas se construye siguiendo la definición dada en la Ecuación (3.1) como, $p^- = (p_1^-, p_2^-) = ((1-p)d_{12}, pd_{21})$. De la Definición 3.3.1 se extrae el método de agregación de pares de evidencias para formar las probabilidades aditivas de las clases C_1 y C_2 como sigue,

$$p^{add} = (p_1^{add}, p_2^{add}) = (\text{máx}\{p - (1-p)d_{12}, 0\}, \text{máx}\{1 - p - pd_{21}, 0\}) .$$

De cara a la asignación final de una instancia x, cuatro escenarios resultan plausibles,

- **Escenario 1.** Si $\frac{p}{1-p} < d_{12}$ y $\frac{1-p}{p} < d_{21}$, entonces $p_1^{add} = 0$ y $p_2^{add} = 0$. Este caso no se satisface nunca ya que $d_{ij} \in [0,1]$ y $\frac{p}{1-p}$ ó $\frac{1-p}{p}$ son mayores que la unidad.
- **Escenario 2.** Si $\frac{p}{1-p} \geq d_{12}$ y $\frac{1-p}{p} < d_{21}$, entonces $p_1^{add} > 0$ y $p_2^{add} = 0$ y por tanto, $C^{add}(x) = C_1$.
- **Escenario 3.** Si $\frac{p}{1-p} < d_{12}$ y $\frac{1-p}{p} \geq d_{21}$, entonces $p_1^{add} = 0$ y $p_2^{add} > 0$, por tanto, $C^{add}(x) = C_2$.
- **Escenario 4.** Si $\frac{p}{1-p} \geq d_{12}$ y $\frac{1-p}{p} \geq d_{21}$, entonces $p_1^{add} = p - (1-p)d_{12} > 0$ y $p_2^{add} = 1 - p - pd_{21} > 0$.

  De este último escenario, se distinguen dos casos,

  $p - (1-p)d_{12} > 1 - p - pd_{21}$ si y solo si $\frac{p}{1-p} > \frac{1+d_{12}}{1+d_{21}}$. Por lo que se tiene que,

$$C^{add}(x) = \begin{cases} C_1 & si \ \frac{p}{1-p} > \frac{1+d_{12}}{1+d_{21}} \\ C_2 & si \ \frac{p}{1-p} < \frac{1+d_{12}}{1+d_{21}} \end{cases}$$

  Con $f(p) = \frac{p}{1-p}$ y $K_D = \frac{1+d_{12}}{1+d_{21}}$. Nótese que $K_D \in [1/2, 2]$.

3.4.2. Comportamiento bajo agregación logística

Supongamos ahora que se dispone de un sistema de clasificación con dos clases en la forma anteriormente descrita, y análogamente se consideran los esquemas de generación de la probabilidad negativa y de agregación de las probabilidades bipolares, respectivamente.

Proposición. En un sistema de clasificación binaria con dos clases $S = \{C_1, C_2\}$ y un clasificador probabilístico C_P, sea x un objeto con $C_P(x) = (p, 1-p)$. Entonces, para todo $p \in (0,1)$, y para todo $j \in \{1,2\}$ existe una matriz D con la cual $C^{log}(x) = C_j$.

Demostración. Dada cierta matriz de disimilitud $D = \begin{pmatrix} 0 & d_{12} \\ d_{21} & 0 \end{pmatrix}$. El vector p^- de probabilidades negativas se construye siguiendo la definición dada en la Ecuación (3.1) como, $p^- = (p_1^-, p_2^-) = ((1-p)d_{12}, pd_{21})$. De la definición 3.3.2 se extrae el método de agregación de pares de evidencias para formar las probabilidades ajustadas de las clases C_1 y C_2 como sigue:

$$p^{log} = (p_1^{log}, p_2^{log}) = \left(1 - e^{-\frac{p}{(1-p)d_{12}}}, 1 - e^{-\frac{1-p}{pd_{21}}}\right).$$

Para la asignación final de clase del clasificador bipolar logístico se puede distinguir entre los siguientes escenarios:

Escenario 1. Si $p = 1$, entonces $p^{log} = (1, 0)$ por definición y, por tanto $C^{log}(x) = C_1$.

Escenario 2. Si $p = 0$, entonces $p^{log} = (0, 1)$ por definición y, por tanto $C^{log}(x) = C_0$.

Escenario 3. Si $p \in (0, 1)$; $d_{12} = 0$ y $d_{21} > 0$, entonces $p^{log} = (1, \alpha)$, con $\alpha \in (0, 1)$ y entonces $C^{log}(x) = C_1$.

Escenario 4. Si $p \in (0, 1)$ y $d_{12}, d_{21} \in (0, 1]$, entonces

$$p^{log} = (p_1^{log}, p_2^{log}) = \left(1 - e^{-\frac{p}{(1-p)d_{12}}}, 1 - e^{-\frac{1-p}{pd_{21}}}\right).$$

Observemos que se satisface la siguiente condición:

$$1 - e^{-\frac{p}{1-p}} > 1 - e^{-\frac{1-p}{p}} \Leftrightarrow \frac{p}{(1-p)d_{12}} > \frac{1-p}{pd_{21}}$$

Por lo que,

$$C^{log}(x) = \begin{cases} C_1 & si \left(\frac{p}{1-p}\right)^2 > \frac{d_{12}}{d_{21}} \\ C_0 & si \left(\frac{p}{1-p}\right)^2 < \frac{d_{12}}{d_{21}} \end{cases}$$

3.4.3. Generación automática de la matriz de disimilitud basada en curvas ROC

El propósito de esta sección es proponer un método automático para la determinación de los valores de las matrices de disimilitud entre clases basado en el clásico juego del movimiento del punto de corte de la probabilidad estimada basada en la curva ROC, de forma que se maximice alguna de las métricas convencionales (accuracy, kappa, etc). El objetivo final es el de extender este método de toma de decisiones finales al caso en el que la variable respuesta a clasificar es de naturaleza nominal con más de dos clases (en el caso de dos clases este método ha de coincidir con el del movimiento del punto de corte de la probabilidad estimada).

Como primera aproximación se realiza un análisis bajo los esquemas *One vs One* o *One vs All* de manera que se convierten los problemas de clasificación multiclase en problemas binarios en los que se considera la decisión dicotómica sobre la pertenencia de las instancias a cierta clase C_i y la pertenencia a otra clase $C_j, j \neq i$ o a la unión de las restantes clases $\bigcup_{j \neq i}(C_j)$, respectivamente. Es decir, bajo el esquema *One vs All*, se descompone un problema de clasificación multiclase de c clases en c problemas de clasificación binaria en los que se enfrenta cada clase al resto de las clases disponibles en el conjunto de datos, mientras que en el esquema *One vs One*, la descomposición se realiza en tantos subproblemas de clasificación binaria como pares de clases, esto es, $\frac{c(c-1)}{2}$, resultando este último método más costoso a nivel computacional.

Una vez obtenido el modelo, se evalúa la clasificación por defecto dada por el punto de corte 0.5 de la probabilidad estimada para la clase de interés y se realiza una búsqueda intensiva de los puntos de corte que maximizan la suma de sensibilidad y especificidad. Con el objetivo de optimizar la búsqueda del punto de corte se genera una rejilla de tantos posibles valores como probabilidades estimadas diferentes y se realiza la clasificación evaluando la métrica considerada y generando una tabla de resultados de la que se selecciona finalmente el α que maximiza la métrica considerada.

Con este punto de corte α se genera la estructura de disimilitud entre clases con la premisa de que para una probabilidad estimada igual a α, la clasificación bipolar no será capaz de tomar una decisión sobre la pertenencia a ninguna de las clases. Para lograr esto, basta con imponer la siguiente condición:

$$\text{máx}\{\alpha - d_{12}(1-\alpha), 0\} = 0 \tag{3.7}$$

Por lo que el valor de disimilitud, para el punto crítico será, $d_{12} = \frac{\alpha}{1-\alpha}$. De manera análoga, este valor para la otra clase se expresa como $d_{21} = \frac{1-\alpha}{\alpha}$, por lo que la matriz tiene la siguiente forma:

$$D_{bin} = \begin{pmatrix} 0 & \frac{\alpha}{1-\alpha} \\ \frac{1-\alpha}{\alpha} & 0 \end{pmatrix}$$

De esta forma, con una sola estructura de disimilitud se consigue replicar el método de movimiento del punto de corte comentado. Por medio de la repetición de este proceso para cada par de clases en el esquema *One vs One* o para cada clase frente al resto *One vs All*, se puede obtener una estructura de disimilitud multiclase basada en la descomposición binaria del problema y la aplicación del paradigma de disimilitudes basadas en el punto de corte óptimo de las curvas ROC. Para ello, teniendo en cuenta un problema de clasificación con c clases, cada elemento d^{ij}_{mult} en la Ecuación 3.8 de la matriz de disimilitud multiclase D_{mult} viene determinado por el correspondiente elemento d^{ij}_{bin} de la matriz disímil en la descomposición binaria D^{ij}_{bin}. De

otro modo, denotando α_{ij} como el punto de corte óptimo de la curva ROC extraído de la clasificación binaria entre las clases i y j, se puede construir D_{mult} como,

$$D_{mult} = \begin{pmatrix} 0 & \frac{\alpha_{12}}{1-\alpha_{12}} & \dots & \frac{\alpha_{1c}}{1-\alpha_{1c}} \\ \frac{1-\alpha_{12}}{\alpha_{12}} & 0 & \dots & . \\ . & \dots & \dots & . \\ \frac{1-\alpha_{1c}}{\alpha_{1c}} & \dots & \dots & 0 \end{pmatrix} \tag{3.8}$$

3.5. Aprendizaje de estructura de disimilitud basada en los datos

El procedimiento de inferencia de la mayoría (si no todos) de los algoritmos de clasificación puede ser interpretado como un proceso de dos etapas en el que la primera de ellas está reservada al entrenamiento del clasificador para producir algún tipo de puntuación *soft* determinando la fuerza de asociación entre un ejemplo y cada clase. La segunda etapa se dedica, en cambio, a la toma final de decisiones o asignación final de clase en base a las puntuaciones alcanzadas en el paso previo.

Desde esta perspectiva, se presenta a lo largo de esta sección la propuesta completa de clasificador en dos etapas señalando, así mismo, las modificaciones fundamentales a introducir en distintas etapas para la creación de las propuestas concretas realizadas en este trabajo.

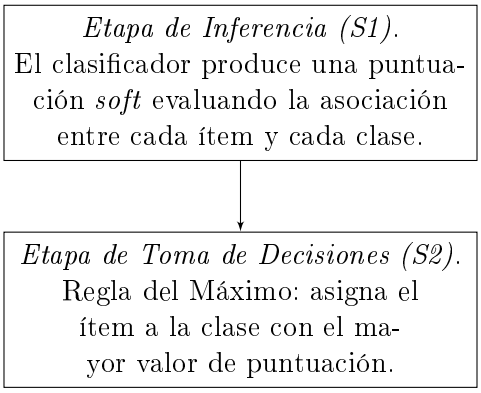


Figura 3.1: Diagrama de flujo de un Clasificador *Soft*.

La Figura 3.1 ilustra la idea recién expuesta, considerando un clasificador general de naturaleza *soft* e interpretando éste como un proceso bietápico. En la primera etapa, el clasificador utiliza el modelo obtenido tras el entrenamiento para desarrollar un mecanismo que permite la evaluación de cada instancia en una contexto *soft*. En la segunda, la clasificación final es llevada

a cabo en base a estas puntuaciones finales por medio de la aplicación de la regla del máximo.

Es evidente que la matriz de disimilitud D juega un papel fundamental en el esquema de clasificación propuesto ya que determina la forma en la que la evidencia negativa es generada desde la evidencia inicial para cada clase. Como consecuencia, la calidad del clasificador nítido resultante así como el efecto de la incorporación de un marco de RBC y el método de agregación o explotación de la información de carácter bipolar presentan una fuerte dependencia de la elección de la matriz D.

La Figura 3.2 muestra el diagrama de flujo de la Etapa de Toma de Decisiones (S2) propuesta, incluyendo el aprendizaje genético de la matriz de disimilitud y su aplicación final al conjunto de prueba. En una situación ideal, la estructura de disimilitud adecuada entre clases debería ser propuesta por expertos en el dominio de aplicación y, por tanto, basado en su conocimiento. Sin embargo, en muchos casos puede ser más práctico obtener esta estructura de clases a través de un proceso de aprendizaje llevado a cabo una vez el clasificador ha alcanzado la evidencia para cada clase.

Cuando este proceso de aprendizaje está conducido por una medida de rendimiento focalizada en la generalización de la precisión de los clasificadores nítidos ajustados, la matriz resultante tiende a permitir la corrección de asignaciones erróneas determinadas por el clasificador base en la muestra de entrenamiento y, con suerte, mejorando también su capacidad predictiva en nuevas instancias de la muestra de prueba. Por tanto, esta propuesta de aprendizaje posibilita que cualquier clasificador pueda beneficiarse de la introducción de una estructura disímil en el conjunto de clases, contribuyendo a una mejor adaptación de la regla de decisión a las características específicas de cada conjunto de datos o contexto de aplicación.

En particular, en este trabajo se propone un proceso de aprendizaje basado en algunas de las metaheurísticas de búsqueda detalladas en la Sección 2.8 del Capítulo 2.

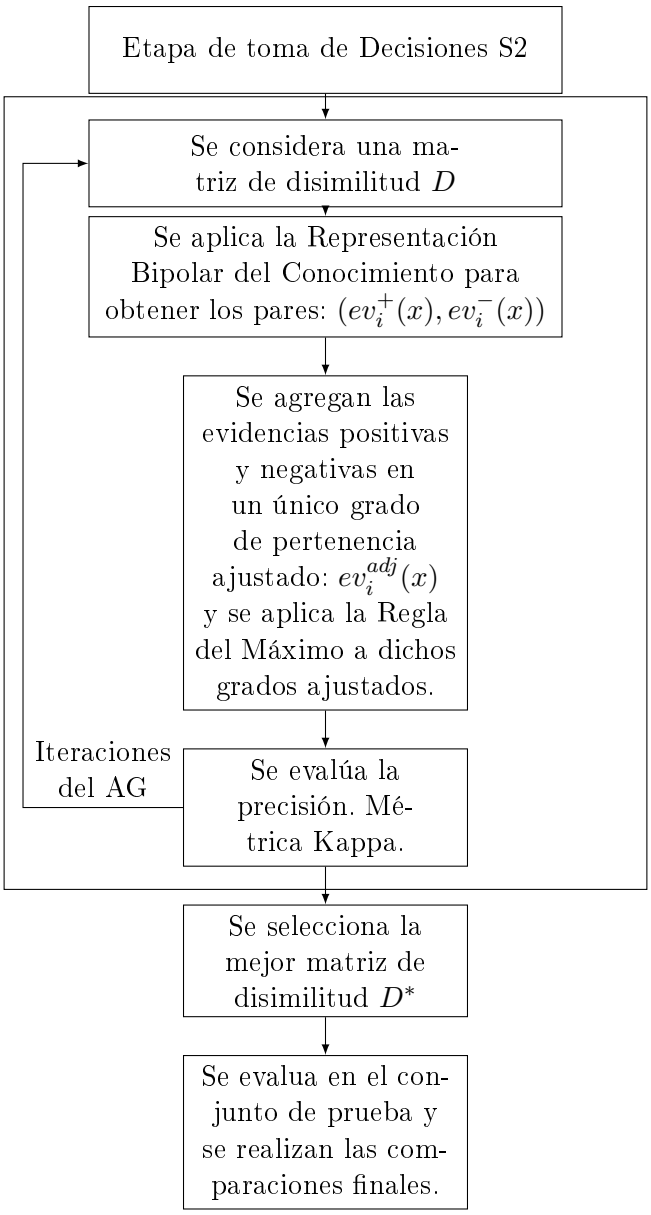


Figura 3.2: Diagrama de Flujo de la Etapa de Toma de Decisiones propuesta (S2)

Capítulo 4

Toma de decisiones bipolar en clasificación supervisada probabilística

Información: el valor recíproco negativo de la probabilidad.

Claude Shannon

Resumen: En este capítulo se proponen y evalúan distintos métodos para la toma de decisiones con representación bipolar del conocimiento en sistemas de clasificación de tipo *probabilístico*. En primer lugar, se presenta en la Sección 4.1, el marco teórico en el contexto de probabilidades, como extensión del margo general introducido en el Capítulo 3. La Sección 4.2 se dedica a la evaluación del marco teórico bajo distintos escenarios. Por un lado se presentan los resultados de las pruebas computacionales orientadas al estudio del comportamiento del esquema propuesto en problemas de clasificación de carácter general en el marco binario (Sección 4.2.2) y multiclase (Sección 4.2.3). Tras la evaluación, se muestran dos aplicaciones prácticas de la explotación de información *soft* dada por clasificadores probabilísticos en el contexto de la accidentalidad vial y la detección de bordes en imágenes (Sección 4.3).

4.1. Marco teórico bipolar-probabilístico

En este capítulo se explora y evalúa el comportamiento de los métodos de Representación Bipolar del Cocimiento (RBC) presentados en este trabajo, en el contexto de los clasificadores probabilísticos definidos en la Sección 2.4.

Recordemos que, de acuerdo a la Ecuación (2.2), todo clasificador probabilístico puede ser visto como una función que asigna a cada instancia x y cada clase C_i con $i \in \{1 \dots c\}$, un valor numérico $p_i \in [0,1]$ que representa la probabilidad de pertenencia del objeto a dicha clase. Así, un clasificador probabilístico devuelve para cada x un vector de probabilidades $C_P(x) = (p_1, \dots, p_c)$ que se encuentra, por definición, sujeto a la restricción $\sum_{i=1}^{c} p_i = 1$.

Cabe señalar que gran parte de los algoritmos de clasificación (Árboles de Clasificación, Regresión Logística, Random Forest, Redes Neuronales Artificiales, etc.) pueden enmarcarse en este grupo de clasificadores probabilísticos si se colecta la información dada por el algoritmo en etapas intermedias previas a la toma de la decisión final.

Es entonces cuando, considerando el vector de probabilidades $C_P(x) = (p_1, \dots, p_c)$ obtenido por el clasificador como el vector de evidencia $ev(x) = (ev_1(x), \dots, ev_c(x))$ presentado en la Sección 3.1 del Capítulo 3, cualquier clasificador probabilístico es susceptible de ser enriquecido mediante la aplicación de una RBC.

Por tanto, es relevante presentar la particularización del marco general de aplicación de la RBC en el contexto de clasificadores de naturaleza *soft* descrito a lo largo del Capítulo 3, en el marco específico de los algoritmos de clasificación supervisada de tipo probabilístico. Así, en adelante se considera que el vector de probabilidades $C_P(x) = (p_1, \dots, p_c)$ dado por el algoritmo, representa la evidencia de carácter positivo en el sentido propuesto en la Sección 3.1, por lo que denotamos $p^+ = (p_1^+, \dots, p_c^+) = (p_1, \dots, p_c)$, que en este caso satisface la restricción $\sum_{i=1}^{c} p_i^+ = 1$.

A continuación, se presentan la extensión de los procedimientos de generación (Sección 4.1.1) y agregación (Sección 4.1.2) de evidencia bipolar detallados en las Secciones 3.2 y 3.3.1, respectivamente.

4.1.1. Obtención de los pares (p^+, p^-)

Es en este punto donde, siguiendo el esquema presentado en la Sección 3.2, se representa por p_i^- la probabilidad de pertenencia de una instancia x a la clase disimil a C_i, entendida como ese constructo abstracto representando el conjunto de clases consideradas disimiles a C_i. Así mismo se denota por $D = (d_{ij})$ la matriz que cuantifica la estructura de disimilitud en el conjunto de clase, donde $d_{i,j}$ se entiende como el grado de disimilitud entre las clases C_i y C_j. Obviamente, se consideran matrices de disimilitud en la forma presentada en la Sección 3.1, es decir satisfaciendo D, $d_{ii} = 0$ para todo $i = 1, \dots, c$ y siendo en general no simétricas.

Tomando en cuenta la estructura de disimilitud y el vector de probabilidades consideradas como evidencia positiva, se construye la evidencia de

carácter negativo p_i^- como,

$$p_i^-(x) = \sum_{j \neq i} d_{ij}\ p_j^+(x) = \sum_{j=1}^{c} d_{ij}\ p_j^+(x) = D_i p^+(x) \tag{4.1}$$

o en forma matricial,

$$p^-(x) = D p^+(x) \tag{4.2}$$

Observación 1. Nótese que, en el caso en que, para una cierta clase C_i, el grado de disimilitud entre C_i y el resto de clases en el conjunto es máximo, esto es, $d_{ir} = 1$ para todo $r \neq i$, entonces $p_i^- = 1 - p_i^+$ que es precisamente la representación del caso clásico en el que la probabilidad negativa es la probabilidad de no pertenencia a la clase C_i. En general y, con mayor frecuencia en estas situaciones donde existe una confusión entre las clases ($d_{ir} < 1$), se tiene que $p_i^- < 1 - p_i^+$.

Observación 2. Es de gran importancia señalar que la probabilidad negativa representa la probabilidad de pertenencia a una clase disimil, por lo que no procede de una distribución de probabilidades, no satisfaciendo necesariamente sus limitaciones.

En este sentido, se presenta el proceso de obtención de los pares de información bipolar en el contexto de los clasificadores de naturaleza probabilística, con la estructura que se muestra en la Figura 4.1.

$$\begin{array}{ccccc} C_P^{bip} : X & \longrightarrow & [0,1]^c & \longrightarrow & [0,1]^c \times [0,1]^c \\ x & \longrightarrow & C_P(x) = (p_1, \ldots, p_c) & \longrightarrow & (p_1^+, \ldots, p_c^+) \times (p_1^-, \ldots, p_c^-) \end{array}$$

Figura 4.1: Esquema de construcción de información bipolar. Primera etapa (E1) de un *Clasificador Bipolar Probabilístico Ajustado* C_P^{bip}

En la Figura 4.1, C_P^{bip} representa un clasificador bipolar probabilístico, en cuya primera etapa (E1), partiendo de un item $x \subset X$ extrae el vector de probabilidades $(p_1, \ldots, p_c) \in [0,1]^c$ y genera los pares de evidencias (en este caso probabilidades) de carácter positivo y negativo $(p_1^+, \ldots, p_c^+) \times (p_1^-, \ldots, p_c^-) \in [0,1]^c \times [0,1]^c$.

Una vez la estructura bipolar pareada ha sido obtenida, uno de entre los muchos métodos de explotación posibles es la utilización de operadores de agregación. En este sentido, se presenta en la Sección 4.1.2, la particularización del marco general de agregación introducido en la Sección 3.3.1 al caso de clasificadores probabilísticos.

4.1.2. Agregación de los pares (p^+, p^-)

Como se ha señalado en la Sección 3.3.1, en el campo de los operadores de agregación existe gran variedad de funciones capaces de representar la información contenida en los pares de probabilidades bipolares (p_i^+, p_i^-), $i \in \{1, \ldots, c\}$ mediante un solo valor numérico ajustado que se puede denotar en este caso como p_i^{adj}. Por tanto, ahora la cuestión clave es la elección de operadores de agregación que unifiquen de forma adecuada las informaciones de carácter positivo y negativo.

$$\begin{array}{ccccc} [0,1]^c \times [0,1]^c & \longrightarrow & [0,1]^c & \longrightarrow & \{C_1, \ldots, C_c\} \\ (p_1^+, \ldots, p_c^+) \times (p_1^-, \ldots, p_c^-) & \longrightarrow & (p_1^{adj}, \ldots, p_c^{adj}) & \longrightarrow & C_{arg\ Max\{p_1^{adj}, \ldots, p_c^{adj}\}} \end{array}$$

Figura 4.2: Esquema de agregación bipolar. Segunda etapa (E2) de un *Clasificador Bipolar Probabilístico Ajustado* C_P^{bip}

En la Figura 4.2, se muestra el diagrama de flujo de la segunda etapa (E2) de lo que se ha llamado *Clasificador Bipolar Probabilístico Ajustado* C_P^{bip}. Este proceso comienza con la consideración de los anteriormente calculados pares de probabilidades $(p_1^+, \ldots, p_c^+) \times (p_1^-, \ldots, p_c^-) \in [0,1]^c \times [0,1]^c$ que, mediante la aplicación de un operador de agregación quedan representados por un único vector de probabilidades ajustadas $(p_1^{adj}, \ldots, p_c^{adj})$. La decisión nítida final es llevada a cabo mediante la conocida regla del máximo, seleccionándose la clase C_i con mayor valor de probabilidad ajustada.

En lo que sigue se definen los dos operadores de agregación ya estudiados en la Sección 3.3.1, ahora en un marco probabilístico. En primer lugar, el operador *Aditivo* probabilístico se define como sigue.

Definición 4.1.1. Sea p_i^+, p_i^- la probabilidades positivas y negativa respectivamente de pertenencia de una instancia a la clase C_i, definimos la Probabilidad Ajustada Aditiva del objeto x a la clase C_i como

$$p_i^{add} = \text{máx}\{0, p_i^+ - p_i^-\}$$

Es preciso interpretar la información dada por el operador aditivo previamente definido, de modo que una probabilidad ajustada aditiva $p_i^{add} > 0$ representa un salto o discrepancia entre la probabilidad de pertenencia a la clase C_i y la probabilidad de pertenencia al constructo definido como *clase disimil* a C_i. Por el contrario, un valor nulo de p_i^{add} es un indicador de la existencia de mayor cantidad de afectos negativos que positivos en la elección de dicha clase.

Del análisis de estructuras de disimilitud presentado en la Sección 3.4, se destacan en este punto las limitaciones de este operador de agregación

aditivo. En efecto, la capacidad de modificación de la clase de pertenencia demostrada por este método de agregación, se ve restringida, al menos en el contexto de la clasificación binaria, a un subconjunto de valores de probabilidad dados por el algoritmo probabilístico base sobre el que se aplica la RBC. Concretamente, el *clasificador bipolar probabilístico aditivo* será capaz de revertir la clasificación inicialmente propuesta por el clasificador probabilístico base si y solo si $p \in [\frac{1}{3}, \frac{2}{3}]$, siendo p la probabilidad dada por el algoritmo base.

Esta restricción es, sin embargo, perfectamente asumible bajo la premisa de una adecuada modelización de las probabilidades llevada a cabo por el clasificador base. En este escenario, los errores de clasificación será cometidos sobre instancias que presentan características comunes a ambas clases, por lo que las probabilidades estimadas deberían ser similares o, en otras palabras, en el entorno del valor $p = 0.5$. Desafortunadamente, en caso de enfrentar un clasificador sesgado o un conjunto de datos desbalanceados, esta restricción supone una limitación real para la toma de decisiones bipolar y el enfoque ha de ser, por ende, ampliado en pos de una mayor representatividad de los distintos escenarios plausibles en clasificación supervisada.

En la siguiente definición se presenta un método alternativo de agregación de las informaciones positiva y negativa en una sola puntuación que hemos llamado *probabilidad ajustada logística* de acuerdo a la notación usualmente utilizada para la función de agregación de tipo exponencial utilizada.

Definición 4.1.2. Sea p_i^+, p_i^- la probabilidades positivas y negativa respectivamente de pertenencia de una instancia a la clase C_i, definimos la Probabilidad Ajustada Logística del objeto x a la clase C_i como

$$p_i^{log} = \begin{cases} 1 - e^{-\frac{p_i^+}{p_i^-}} & si\ p_i^- > 0 \\ 1 & en\ otro\ caso \end{cases}$$

Como se ha destacado en la Sección 3.3.1 la agregación logística está basada en el ratio entre informaciones positiva y negativa, ajustando éste al rango $[0, 1]$ a través de una transformación logística, por lo que este esquema permite un comportamiento de alguna forma más flexible de los grados ajustados.

Para finalizar el proceso llevado a cabo por el *Clasificador Bipolar Probabilístico* C_P^{bip}, tras la aplicación de alguna de las agregaciones anteriormente propuestas y las probabilidades bipolares ajustadas $p_i^{adj}(x)$ han sido obtenidas para cada clase (ya sea $p_i^{adj}(x) = p_i^{add}(x)$ o $p_i^{adj}(x) = p_i^{log}(x)$), la decisión nítida final en la clasificación del item x se lleva a cabo por aplicación de la regla del máximo sobre estos grados bipolares ajustados como se observa en el esquema de la Figura 4.2.

Para concluir la sección, y como concepto general que engloba el proceso descrito, se define formalmente un *Clasificador Bipolar Probabilístico*

Ajustado C_P^{bip} a continuación.

Definición 4.1.3. Dado un clasificador probabilístico $C_P : X \longrightarrow [0,1]^c$ en un universo de discurso X y dada una matriz de disimilitud D que permita crear la información bipolar $(p^+, p^-) = \big((p_1^+, p_1^-), \ldots, (p_c^+, p_c^-)\big)$ para cada instancia x a partir de la información dada por $C_P(x) = (p_1, \ldots, p_c)$, se puede definir el *Clasificador Bipolar Probabilístico Ajustado* como

$$C_P^{bip}(x) = C_h \text{ si y solo si } p_h^{adj} = \text{máx}\{p_j^{adj}; j = 1 \ldots c\}.$$

donde p^{adj} es la probabilidad ajustada resultado de la agregación de p^+ y p^-.

En particular, se denotarán en adelante como C_P^{bipAdd} y C_P^{bipLog} los clasificadores bipolares probabilísticos ajustados aditivo y logístico respectivamente, construidos a partir de la información dada por cierto clasificador C.

Observación 3. Cabe destacar que, el marco general de aplicación de la RBC no se restringe al uso de operadores de agregación para la explotación de la evidencia (probabilidades en este caso) bipolar (ver Sección 3.3.2). Por el contrario, distintos tipos de esquemas de explotación podrían ser considerados. En este sentido, se extiende la definición anterior teniendo en cuenta que ahora $C_P^{bip}(x) = C_r$ tal que $\Phi(p^+, p^-) = C_r$, siendo Φ la función de explotación general considerada.

4.2. Resultados Experimentales

Esta sección se dedica por completo a la presentación de los resultados experimentales relativos a las aportaciones de este capítulo. En la primera parte se evalúan los métodos propuestos en un nivel general, esto es, a través de una configuración experimental apropiada para la comparación de algoritmos en el ámbito de la MDD. Se distinguen dos casos diferenciados por el tipo de problema de clasificación considerado: binario o multiclase. En la Sección 4.2.2 se detallan los resultados obtenidos en la contribución [80]. El caso de los clasificadores multiclase es abordado en la Sección 4.2.3, en la que los resultados se extraen directamente de la contribución [82]. En cada caso se especifican los detalles de la configuración experimental escogida así como las mejoras obtenidas y, finalmente se apuntan las conclusiones más relevantes. En una segunda parte, se muestran aplicaciones a realidades particulares definidas por conjuntos de datos de carácter específico, por un lado en el campo de la siniestralidad vial (Sección 4.3.1) y por otro lado la detección de bordes en el procesamiento de imágenes (Sección 4.3.2).

4.2.1. Configuración Experimental

Se detallan en esta sección los pormenores de la configuración experimental seguida para la evaluación del comportamiento de los métodos propuestos en el marco general de comparación de algoritmos en MDD. Debido al similar carácter de esta configuración en los casos binario y multi-clase, se decide presentar la misma en una única sección. Por tanto, se dan detalles al respecto de los conjunto de datos utilizados y su fuente, así como de la configuración experimental, de los parámetros del Algoritmo Genético utilizado y del análisis estadístico aplicado.

4.2.1.1. Configuración General

Para la evaluación del comportamiento de las propuestas descritas en este capítulo en los marcos binario y multi-clase, se seleccionan bancos de pruebas compuestos por 10 y 18 conjuntos de datos, respectivamente, extraídos del repositorio de conjunto de datos KEEL [75], que están disponibles públicamente en el sitio web `www.keel.es/datasets.php` incluyendo información general sobre ellos, particiones para la validación de los resultados experimentales y utilidades adicionales.

Se considera un modelo de validación cruzada de 5-*folds* para llevar a cabo los diferentes experimentos. Es decir, los conjuntos de datos considerados han sido divididos en 5 particiones aleatorias de datos, cada una con 20 % de patrones, y empleamos una combinación de 4 de ellos (80 %) para entrenar el sistema y el restante para probarlo.
Los resultados para cada clasificador en cada experimento se obtendrán siguiendo un esquema de validación cruzada *5-folds* para cada conjunto de datos. En cada *fold*, es decir, para cada conjunto de entrenamiento , la configuración paramétrica óptima del clasificador base se aproxima utilizando una cuadrícula P en el espacio de parámetros propios del algoritmo considerado. Para evaluar el rendimiento de cada configuración paramétrica específica $p \in P$, se generan 25 muestras *boostrap* del conjunto de entrenamiento , de tal manera que los clasificadores básicos se ajustan a cada una de estas muestras y luego se prueban en un esquema de muestreo *leave-one-out* (entrenando el algoritmo con todas las instancias excepto una de ellas y probando el rendimiento en esta) mediante la utilización del estadístico kappa.

En cada *fold*, el proceso de aprendizaje genético de los parámetros de la bipolaridad se lleva a cabo a partir de los vectores de probabilidades estimadas $p(x)$ de los elementos x en la muestra de entrenamiento de la forma mostrada en 3.2.

Las medidas de rendimiento en entrenamiento y prueba de cada uno de los clasificadores para cada conjunto de datos considerado en cada experimento, se calculan finalmente promediando los índices de precisión en los conjuntos de entrenamiento y prueba de las $F = 5$ *folds*.

4.2.1.2. Datos para clasificación binaria

La Tabla 4.1 resume las propiedades de los 10 conjuntos de datos seleccionados para el caso binario, mostrando para cada conjunto de datos el número de ejemplos (#Ex.), el número de atributos (#Atts.), el tipo de atributo (Real/Entero/Nominal) y la relación de desequilibrio (#IR) una vez que el conjunto de datos se ha transformado en un problema de clasificación binaria. La relación de desequilibrio (IR), dado como cociente entre el número de instancias de las clases con mayor y menor representación en el conjunto de datos [34].

Id.	**Data-set**	**#Ex.**	**#Atts.**	**(R/I/N)**	**#IR**
aut	Autos	159	25	(15/0/10)	2.58
con	Contraceptive	1473	9	(6/0/3)	3.43
fla	flare	1066	25	(15/0/10)	2.58
eco	ecoli	336	7	(7/0/0)	3.34
gla	Glass	214	9	(9/0/0)	2.05
lin	Lymphography	148	18	(3/0/15)	1.43
shu	Shuttle	2175	9	(0/9/10)	5.44
thy	Thyroid	720	21	(6/0/15)	18.1
yea	Yeast	1484	8	(8/0/0)	5.1
car	Car	159	25	(15/0/10)	3.5

Tabla 4.1: Descripción de los conjunto de datos empleados. Clasificación probabilística binaria.

Cabe destacar que el conjunto de datos *Thyroid* presenta un alto desbalanceo de clases (entendido como una gran diferencia entre las clases con mayor y menor representación) con $IR = 18.1$, debido al desbalanceo inherente a su distribución original multiclase. Se puede considerar este hecho como una oportunidad para evaluar el comportamiento de los métodos bipolares en conjuntos de datos con alto desbalanceo.

4.2.1.3. Datos para clasificación multi-clase

La Tabla 4.2 resume las propiedades de cada uno de los 18 conjuntos de datos de clasificación multi-clase, donde se muestran el número de ejemplos (#Ex.), el número de atributos (#Atts.) y tipo (Real/Entero/Nominal). En esta ocasión, para transformar conjuntos de datos originales en conjuntos de datos de tres clases, se toma como clase C_0 y C_1 los originales más cercanos a 20 % de instancias y como clase C_2 la unión de las clases restantes.

Id.	Data-set	#Ex.	#Atts.	(R/I/N)
Aut	Autos	159	25	(15/0/10)
Car	Car	159	25	(15/0/10)
Wnq	Winequality-red	1599	11	(11/0/0)
Pen	Penbased	10992	16	(0/16/0)
Pag	Page-blocks	5472	10	(4/6/0)
Der	Dermathology	366	34	(0/34/0)
Eco	ecoli	336	7	(7/0/0)
Fla	flare	1066	25	(15/0/10)
Gla	Glass	214	9	(9/0/0)
Shu	Shuttle	2175	9	(0/9/10)
Yea	Yeast	1484	8	(8/0/0)
Lin	Lymphography	148	18	(3/0/15)
Bal	Balance	625	4	(4/0/0)
Win	Wine	178	13	(13/0/0)
Nty	Newthyroid	215	5	(4/1/0)
Hay	Hayes-Roth	160	4	(0/4/0)
Con	Contraceptive	1473	9	(6/0/3)
Thy	Thyroid	720	21	(6/0/15)

Tabla 4.2: Descripción de los conjuntos de datos utilizados para la evaluación de propuestas en contexto multi-clase.

4.2.1.4. Algoritmo Genético

En cuanto a las especificaciones del algoritmo genético utilizado, se trata del AG implementado en el paquete *genalg* de R [84] con la siguiente configuración paramétrica:

- Tamaño de la población: 50 individuos
- Número de iteraciones: 20
- *Mutation chance* (la posibilidad de que un gen en el cromosoma mute): 0.01
- *Elitism* (el número de cromosomas que se mantienen en la próxima generación): El 20 % del tamaño de la población.

4.2.1.5. Análisis Estadístico

De cara a la evaluación de resultados, un completo análisis estadístico es llevado a cabo siguiendo las recomendaciones dadas en [32, 33] y considerando el estadístico kappa [15] como métrica para la evaluación.
En todos los análisis presentados a lo largo de este trabajo se utilizan algunas técnicas de validación de hipótesis para dar soporte estadístico al análisis de

los resultados. Se consideran pruebas no paramétricas ya que las condiciones iniciales que garantizan la fiabilidad de las pruebas paramétricas pueden no ser satisfechas, lo que implica que el análisis estadístico puede perder credibilidad cuando se usan pruebas paramétricas en este contexto [18].

Específicamente, utilizamos los test de rangos alineados de Friedman, que se recomienda en la literatura [18, 32, 33] para detectar diferencias estadísticas entre un grupo de resultados. Finalmente, la prueba post hoc de Holm [42] se ha utilizado para encontrar los algoritmos que rechazan la hipótesis de igualdad con respecto a un método de control seleccionado. Este procedimiento post-hoc permite saber si una hipótesis de comparación de medias podría rechazarse a un nivel específico de significación *alpha*. Además, calculamos el $p - valor$ ajustado (APV) para tener en cuenta que se realizan múltiples pruebas. De esta manera, podemos comparar directamente el APV con respecto al nivel de significación α para poder decidir si rechazar o no la hipótesis nula. Se utiliza, así mismo, el test de Wilcoxon [83] para realizar las comparaciones por pares de métodos.

Una completa descripción de estas pruebas junto con muchas consideraciones y recomendaciones, e incluso el software utilizado para ejecutar este análisis se encuentra disponible en el sitio web `sci2s.ugr.es/sicidim`.

En resumen, para evaluar la idoneidad de nuestra propuesta bipolar, queremos mostrar empíricamente si esta metodología mejora los resultados del clasificador de base, esto es, sin el paso de ajuste bipolar. En este sentido, para cada clasificador base tenemos dos funciones de agregación. Por lo tanto, el objetivo principal es comparar, para cada clasificador, el rendimiento alcanzado por los tres enfoques: Clasificador sin ajuste, Clasificador + ajuste bipolar aditivo y Clasificador + ajuste bipolar logístico.

El estudio experimental se ha obtenido utilizando R Software. Específicamente, usamos el paquete *caret* [48] para el entrenamiento de clasificadores, ajustándolos a través de los paquetes clásicos subyacentes *random forest* y *nnet*, y finalmente el paquete *genalg* [84] para evaluar la AG.

4.2.2. Clasificación binaria

Esta sección tiene como objetivo analizar el comportamiento de la RBC cuando se aplica sobre reconocidos clasificadores como CART [8], Random Forest (RF) [9] y Redes neuronales (ANN) [64, 78]. Para describir el contenido de la contribución [80], se muestran y valoran los resultados obtenidos y se realiza posteriormente un análisis estadístico de los mismos. Finalmente se discuten los resultados y se extraen las principales conclusiones de este trabajo.

4.2.2.1. Resultados

Siguiendo las especificaciones detalladas anteriormente, se realizan las pruebas experimentales sobre los conjunto de datos considerados y se agrupan los resultados, para cada algoritmo base, en pares de entrenamiento y prueba, donde se destaca en **negrita** el mejor resultado global para cada conjunto de datos de tets.

Se observa, a partir de los resultados de las Tablas 4.3, 4.4 y 4.5 el buen comportamiento general del método de ajuste bipolar, al menos en uno de los enfoques de agregación, ya que mejora el rendimiento de los algoritmos base iniciales.

	CART		CARTbipAdd		CARTbipLog	
	Tr.	*Tst.*	*Tr.*	*Tst.*	*Tr.*	*Tst.*
aut	.730	.738	.730	.738	.730	.738
car	.793	.724	.793	**.733**	.793	**.733**
con	.328	.241	.337	.295	.339	**.298**
eco	.714	.598	.714	.598	.714	.598
fla	.584	.529	.595	**.530**	.595	**.530**
gla	.714	.531	.714	.531	.714	.531
lin	.659	.514	.659	.514	.659	.514
shu	.995	.993	.995	.993	.995	.993
thy	.890	.823	.890	.823	.890	.823
yea	.521	.436	.524	**.447**	.524	**.447**
Mean	.693	.613	.695	.620	.695	**.621**

Tabla 4.3: Resultados en los conjuntos de entrenamiento (*Tr.*) y prueba (*Tst.*) alcanzados por las propuestas bipolares genéricas aplicadas sobre el algoritmo CART.

Cuando el método bipolar es aplicado al clasificador CART, Tabla 4.3, se alcanzan mejoras en 4 de cada 10 conjuntos de datos y ambas agregaciones presentan un comportamiento muy similar.

Contrariamente a lo acontecido con CART, para el clasificador de RF está claro que la agregación aditiva de información tanto positiva como negativa supera al enfoque logístico, obteniendo mejoras en 6 de los 10 conjunto de datos. Como se adelantó en el apartado de configuración experimental (Sección 4.2.1), en ocasiones se puede dar la situación límite de perfecto ajuste del clasificador base en el conjunto de train con claro sobreajuste (ya que en el conjunto de prueba el valor de kappa desciende drásticamente). Este es claramente el caso del RF. Enfrentando esta situación según lo pre establecido, se permite al AG encontrar una estructura de disimilitud de entre aquellas que no producen cambios en la clasificación en entrenamiento pero susceptibles de poder hacerlo en el conjunto de prueba. Así, se producen

	RF		RFbipAdd		RFbipLog	
	Tr.	*Tst.*	*Tr.*	*Tst.*	*Tr.*	*Tst.*
aut	1	.760	1	**.787**	1	.779
car	1	.796	1	**.805**	1	.804
con	.868	.235	.884	.252	.884	**.259**
eco	1	.578	1	**.581**	1	.580
fla	.646	.549	.672	.571	.673	**.575**
gla	1	.666	1	**.701**	1	.691
lin	.986	.684	1	**.702**	1	.677
shu	1	**.996**	1	.995	1	.995
thy	1	.885	1	**.899**	1	.883
yea	1	**.503**	1	.495	1	.491
Mean	.950	.665	.956	**.679**	.956	.673

Tabla 4.4: Resultados en los conjuntos de entrenamiento (*Tr.*) y prueba (*Tst.*) alcanzados por las propuestas bipolares genéricas aplicadas sobre el algoritmo RF.

algunas mejoras bajo este paradigma, como en el caso de los archivos *autos, car* y *glass*.

	ANN		ANNbipAdd		ANNbipLog	
	Tr.	*Tst.*	*Tr.*	*Tst.*	*Tr.*	*Tst.*
aut	.686	.568	.703	**.610**	.703	**.610**
car	.992	.954	.995	.948	.995	**.949**
con	.254	.235	.382	.271	.388	**.285**
eco	.662	.555	.693	**.581**	.692	**.581**
fla	.597	.589	.626	**.595**	.633	.589
gla	.568	.442	.647	.476	.645	**.480**
lin	.895	**.759**	.905	.744	.906	.745
shu	.962	.955	.980	**.966**	.981	.965
thy	.856	.628	.933	.813	.941	**.817**
yea	.518	.467	.558	**.529**	.562	.528
Mean	.699	.615	.742	.653	.745	**.655**

Tabla 4.5: Resultados en los conjuntos de entrenamiento (*Tr.*) y prueba (*Tst.*) alcanzados por las propuestas bipolares genéricas aplicadas sobre el algoritmo ANN.

En el caso de ANN, como evidencian los resultados de la Tabla 4.5, en términos generales, la agregación logística parece alcanzar mejores resultados que la aditiva.

4.2.2.2. Análisis estadístico

Con el fin de detectar diferencias significativas entre los resultados de los diferentes enfoques, se considera el marco de comparación estadística de modelos en clasificación supervisada propuesto en [18, 33]. En primer lugar llevamos a cabo el test de rangos alineados de Friedman que es adecuado para comparaciones múltiples. Esta prueba obtiene un p-valor bajo para los tres algoritmos, lo que implica que hay diferencias significativas entre los resultados. Por este motivo, se considera un *test* post-hoc para comparar nuestra metodología con los enfoques restantes. Específicamente, se aplica un test Holm utilizando el mejor enfoque (el que tiene rango más bajo) como método de control y calculando el p-valor ajustado (APV), indicando con * los valores significativos a nivel de confianza del 95 %. Este análisis estadístico, que se muestra en la Tabla 4.6, refleja que el método bipolar supera al clasificador base con un alto nivel de confianza para todos los algoritmos considerados.

Algoritmo	Rango CART	Rango RF	Rango ANN
Ref"	21.9	22.2	23.4
"BipAdd"	12.35	9.45	11.9
"BipLog"	12.25	14.85	11.2
p-val	.0271	.0255	.0295
Holm APV	**.0284***	**.0024***	**.0038***

Tabla 4.6: Rangos promedio de los algoritmos (Aligned Friedman), p-valores asociados y p-valor Ajustado de Holm para cada clasificador base.

El análisis estadístico llevado mediante un test de Wilcoxon, ver Tabla 4.7, refleja claramente la superioridad de esta nueva metodología con un p-valor aceptable. En el caso de la RF, solo la agregación aditiva podría considerarse mejor que la referencia en términos estadísticos.

Comparación	R^+	R^-	Ex. p-val	Asint. p-val
CARTbipAdd vs. CART	44.5	10.5	.094	.074
CARTbipLog vs. CART	44.5	10.5	.094	.074
RFbipAdd vs. RF	51.0	4.0	.013	.014
ANNbipAdd vs. ANN	50.0	5.0	.019	.019
ANNbipLog vs. ANN	48.0	7.0	.037	.032

Tabla 4.7: Test de Wilcoxon para comparar los métodos de ajuste bipolar (R^+) frente al clasificador base (R^-).

4.2.2.3. Conclusiones

En la contribución que se desprende de esta sección de resultados en clasificación binaria [80] se presenta un clasificador basado en la representación

del conocimiento bipolar, una propuesta para ajustar la clasificación dada por cualquier algoritmo de clasificación.

Para ello, se desarrolla un método basado en AG para encontrar la estructura de disimilitud óptima entre las clases así como dos nuevos enfoques para la agregación de informaciones de carácter positivo y negativo. A lo largo del estudio experimental, hemos aprendido varias lecciones:

- El método bipolar permite mejorar los resultados de los tres algoritmos base de MDD considerados en este trabajo.
- Las agregaciones aditiva y logística superan en rendimiento los resultados del clasificador de base en el caso de CART y ANN, y solo el método de agregación aditiva mejora el comportamiento de RF en términos estadísticos.
- Comparando los dos métodos de agregación, no existe un claro ganador, de hecho, es altamente dependiente del algoritmo base considerado, así como del conjunto de datos de aplicación.

Estos resultados permiten concluir que esta nueva metodología es una solución adecuada para enfrentar los problemas de clasificación binaria al incorporar el marco de representación de conocimiento bipolar flexible a la información proporcionada por cualquier algoritmo de naturaleza probabilística.

Una vez comprobado el buen comportamiento del método bipolar probabilístico en problemas de clasificación binarios, se pretende extender el análisis al marco multiclase con el fin de evaluar el comportamiento del paradima de aprendizaje de disimilitudes cuando existen mas de dos clases implicadas.

4.2.3. Clasificación multiclase

Esta sección está dedicada a presentar la evaluación, recogida en la contribución [82], del rendimiento de los enfoques de representación de conocimiento bipolar basados en disimilitud (aditivos y logísticos) cuando se aplican en reconocidos clasificadores como Random Forest (RF) [9] y Redes Neuronales (ANN) [64, 78] en un contexto de problemas de clasificación con tres clases en la variable de interés para la clasificación. El objetivo es valorar la extensión del método bipolar en entorno probabilístico, que fue probado en la anterior sección para problemas binarios, esta vez, en el contexto de clasificación multiclase.

En las siguientes secciones se presenta, en primer lugar, la configuración experimental establecida, seguida de los resultados obtenidos y el análisis estadístico correspondiente para, finalmente dar una breve discusión de resultados y las conclusiones destacables.

4.2.3.1. Resultados

Esta sección tiene como objetivo presentar los resultados del experimento computacional descrito anteriormente y recogido en [82], llevada a cabo para estudiar la capacidad de mejora de nuestros clasificadores ajustados bipolares con respecto al clasificador de base de referencia al que se aplica el método de ajuste de decisión final propuesto.

Los resultados se agrupan, para cada algoritmo base, en pares para entrenamiento y prueba, donde el mejor resultado global para cada conjunto de datos considerado es señalado en **negrita**. No destacando ninguno en caso de empate.

Se observa en los resultados de las tablas 4.8 y 4.9, el buen comportamiento general del método de ajuste bipolar, al menos con respecto a uno de los métodos de ajuste bipolar, ya que permite la mejora en el rendimiento de los algoritmos de referencia.

	RF					
	Ref		bipAdd		bipLog	
	Tr.	*Tst.*	*Tr.*	*Tst.*	*Tr.*	*Tst.*
Aut	1	.716	1	**.719**	1	.706
Car	.996	**.867**	1	.854	1	.857
Wnq	1	.515	1	.489	1	**.525**
Pen	1	**.903**	1	.895	1	.892
Pag	1	.831	1	**.832**	1	**.832**
Der	1	**.995**	1	.993	1	.992
Eco	1	.758	1	**.775**	1	.764
Fla	.796	.783	.805	**.787**	.807	.784
Gla	1	.672	1	.658	1	**.677**
Shu	1	**.996**	1	**.996**	1	.995
Yea	1	.377	1	.366	1	**.378**
Lin	.981	.672	.996	.675	.996	**.710**
Bal	.612	**.556**	.615	.523	.617	.513
Win	1	**.979**	1	.954	1	.973
Nty	1	**.935**	1	.912	1	.895
Hay	.885	.703	.886	**.715**	.886	**.715**
Con	.788	.280	.807	**.286**	.807	.279
Thy	1	.895	1	**.897**	1	.891
Mean	.948	**.746**	.950	.740	.951	.743

Tabla 4.8: Resultados en los conjuntos de entrenamiento (*Tr.*) y prueba (*Tst.*) alcanzados por los clasificadores genéticos bipolares aplicados sobre el algoritmo base RF.

En cuanto a la aplicación del método bipolar probabilístico sobre el al-

goritmo RF, la Tabla 4.8 refleja los resultados obtenidos, pudiendo extraer las siguientes características sobre su comportamiento.

- No existe mejora en kappa cuando se compara el modelo bipolar aditivo contra referencia.
- El clasificador bipolar aditivo supera la clasificación de los enfoques restantes en 7 de los 18 conjuntos de datos y el logístico lo hace en 6 de ellos.
- La referencia gana en 6 de los 18 conjuntos de datos.
- Existe un empate entre el enfoque bipolar aditivo y la referencia en el conjunto de datos de Shuttle.

Por lo tanto, se puede ver que se alcanzan mejoras o empates en 12 de los 18 conjuntos de datos cuando se considera alguna de las dos agregaciones propuestas. Es importante tener en cuenta el comportamiento variable del método bipolar aditivo en este caso. A pesar de ser el método ganador en número de conjuntos de datos, podemos ver que su media no es la mejor debido al menor valor kappa obtenido en varios de los conjuntos de datos restantes.

Teniendo en cuenta el clasificador ANN, el método bipolar alcanza los resultados que se muestran en la Tabla 4.9 que podría interpretarse de la siguiente manera:

- Hay una mejora en kappa de .004 al comparar el modelo logístico bipolar frente a referencia, siendo de .002 en el caso del aditivo.
- Los clasificadores bipolares aditivos y logísticos superan en rendimiento la clasificación de las restantes aproximaciones en 7 y 9 de los 18 conjuntos de datos, respectivamente.
- el algoritmo de referencia gana en 4 de los 18 conjuntos de datos.
- Se producen dos empates en estos resultados.

En general, se obtienen mejoras o empates en 14 de los 18 conjuntos de datos al comparar los enfoques bipolares con la referencia. Con el fin de detectar diferencias significativas entre los resultados de los diferentes enfoques, llevamos a cabo el test de rangos alineado de Friedman. Esta prueba obtiene un p-valor bajo para los tres algoritmos, lo que implica que existen diferencias significativas entre los resultados proporcionados por cada método.

	ANN					
	Ref		bipAdd		bipLog	
	Tr.	*Tst.*	*Tr.*	*Tst.*	*Tr.*	*Tst.*
Aut	.504	.382	.533	**.385**	.532	**.385**
Car	1	.997	1	.997	1	.997
Wnq	.359	.341	.399	**.356**	.399	**.356**
Pen	.954	.855	.964	**.866**	.966	.856
Pag	.853	.753	.874	.755	.887	**.774**
Der	1	.987	1	**.991**	1	**.991**
Eco	.753	**.697**	.779	.680	.777	.688
Fla	.785	**.788**	.794	.782	.795	.777
Gla	.660	.507	.688	**.517**	.687	.513
Shu	.991	.976	.993	**.977**	.994	**.977**
Yea	.440	.360	.473	.379	.473	**.381**
Lin	.896	.667	.922	.671	.925	**.678**
Bal	.600	**.586**	.603	.562	.603	.563
Win	.945	.911	.959	**.915**	.960	.901
Nty	.986	.957	.995	.957	.997	.957
Hay	.811	**.615**	.850	.600	.845	.588
Con	.356	.334	.383	.336	.383	**.338**
Thy	.859	.738	.904	.770	.925	**.803**
Mean	.764	.692	.784	.694	.786	**.696**

Tabla 4.9: Resultados en los conjuntos de entrenamiento (*Tr.*) y prueba (*Tst.*) alcanzados por los clasificadores genéticos bipolares aplicados sobre el algoritmo base ANN.

4.2.3.2. Análisis estadístico

En cuanto a la configuración del aparato estadístico utilizado para evaluar la significación de los resultados obtenidos, de nuevo se recurre al esquema de pruebas no paramétricas seguido en la Sección 4.2.2 y cuya completa descripción así como diversas recomendaciones e incluso el software para ejecutar dichas pruebas se encuentra en `sci2s.ugr.es/sicidm`.

Se aplica, entonces, una prueba post hoc para comparar esta metodología con los enfoques restantes. Específicamente, una prueba Holm se aplica utilizando el mejor enfoque (el que tiene el menor valor de rango) como método de control y calculando el p-valor ajustado (APV) para el método con mayor rango.

Obviamente, sería deseable que la referencia alcanzase el rango más alto o, al menos, no el más bajo, ya que generalmente se asocia con peores resultados.

La Tabla 4.10, refleja la existencia de diferencias estadísticamente signi-

Algoritmo	Rango RF	Rango ANN
Ref"	22.22	31.5
"BipAdd"	31.83	26.44
"BipLog"	28.44	24.55
p-val	.00097	.000913
APV	**.1336**	**.371**

Tabla 4.10: Rangos promedio de los algoritmos (Aligned Friedman), p-valores asociados y p-valor Ajustado de Holm para cada algoritmo.

ficativas entre los tres clasificadores para los algoritmos de RF y ANN. Sin embargo, en caso de RF estas diferencias del análisis estadístico deben ser interpretadas con cautela debido al menor valor de rango obtenido por el algoritmo de referencia. De hecho, la referencia (RF sin aplicar ningún enfoque bipolar) parece alcanzar los mejores resultados con respecto al test de Friedman, a pesar de presentar mejor rendimiento en solo 4 de los 18 conjuntos de datos. Por lo tanto, no hay evidencia estadística de la superioridad de ninguno de los métodos comparados en el caso de RF.

Con respecto al clasificador ANN base, la Tabla 4.10 muestra la superioridad de ambos enfoques bipolares en los valores de clasificación, sin embargo, la prueba post-hoc de Holm refleja que no hay evidencia suficiente para garantizar que ambos enfoques bipolares superen la referencia.

Comparación	R^+	R^-	p-val
RFbipAdd vs. RFRef	115.0	56.0	.1913
RFbipLog vs. RFRef	100.0	71.0	.5135
ANNbipAdd vs. ANNRef	100.0	53.0	.2559
ANNbipLog vs. ANNRef	95.0	58.0	.3684

Tabla 4.11: Test de Wilcoxon para comparar los métodos bipolares (R^+) frente al clasificador base (R^-).

El análisis estadístico de las comparaciones por pares de métodos, que se realiza mediante una prueba de Wilcoxon, Tabla 4.11, refleja la débil superioridad de la metodología propuesta cuando se aplica a los algoritmos de RF y ANN con p-valor relativamente bajo en caso del modelo bipolar aditivo. Nuevamente, la aplicación de la metodología en el algoritmo de RF y ANN no alcanza mejoras significativas.

4.2.3.3. Conclusiones

En este trabajo se estudia y evalúa el comportamiento del paradigma de Representación Bipolar de Conocimiento (RBC) en el entorno de clasificadores probabilísticos en problemas multiclase. La introducción de estructuras

de disimilitud en clasificación, puede permitir una mayor adaptación de los clasificadores a cada contexto de aplicación específico, en el que las clases adquieren una semántica particular, lo que también permite una mejora en el rendimiento del clasificador.

En este sentido, el enfoque propuesto puede entenderse como un método general de procesamiento "a posteriori" para ajustar la regla del máximo que generalmente se aplica para tomar la decisión sobre la asignación de clase de cada elemento a clasificar.

Para estudiar la viabilidad del enfoque propuesto, y en particular para señalar que es susceptible de aplicación a cualquier clasificador *soft* independientemente de como éste sea entrenado, se propone la aplicación a dos de los clasificadores supervisados más poderosos, Random Forest (RF) y Redes Neuronales Artificiales (ANN). A través de una rigurosa experimentación, se analiza si los enfoques bipolares aditivos y logísticos propuestos permitieron una mejora estadísticamente significativa de los clasificadores probabilísticos básicos en el contexto multiclase.

A lo largo de este estudio experimental, se extraen las siguientes conclusiones:

- El marco bipolar mejoró los resultados de los dos algoritmos de clasificación base considerados en este trabajo en cuanto a número de conjuntos de datos con mayor índice kappa.
- Los métodos de ajuste logístico y aditivo no superaron significativamente los resultados del clasificador base. Sin embargo, alcanzaron p-valores relativamente bajos en la prueba de Wilcoxon, especialmente el aditivo.
- Comparando los clasificadores aditivo y logístico, encontramos que no hay un ganador claro. De hecho, esta cuestión parece depender de alguna manera del algoritmo base considerado, así como del conjunto de datos de la aplicación.

Estos resultados nos llevan a la conclusión de que el enfoque propuesto podría proporcionar una solución adecuada para enfrentar los problemas de clasificación de tres clases y mejorar la regla de decisión que determina cómo se explota la información flexible intermedia recopilada por muchos clasificadores. Para ello distintas configuraciones experimentales merecen ser consideradas.

Es relevante destacar que, debido a la configuración experimental seguida, en la que se entrenan los algoritmos probabilísticos base con el objetivo de encontrar la configuración paramétrica óptima, dificulta en cierto modo el aprendizaje de una estructura disímil adecuada para mejorar la capacidad de generalización a nuevos datos. Un caso extremo es el comentado anteriormente, cuando el clasificador base obtiene un ajuste perfecto a los datos de

entrenamiento . De cara a manejar estas situaciones difíciles, se considera un paradigma de aprendizaje en el que se permite al AG encontrar una estructura de disimilitud susceptible de producir mejoras en el conjunto de prueba, asumiendo con gallardía en este punto el riesgo de obtener peores resultados finales.

4.3. Aplicaciones

En esta sección se muestran dos aplicaciones prácticas de la explotación de información *soft* dada por clasificadores probabilísticos en el contexto de la accidentalidad vial (Sección 4.3.1) y la detección de bordes en imágenes (Sección 4.3.2).

4.3.1. Datos de la DGT

Esta sección se reserva para mostrar la aplicación al problema real que supuso el germen para el desarrollo de esta tesis, los datos de siniestralidad vial. Este trabajo parcial se encuentra recogido en la contribución [79]. Concretamente, la fuente de información del presente estudio la constituye la base de datos de accidentes ocurridos en el año 2012 en España, proporcionada por la DGT. Dicha base de datos está integrada por 83.115 registros que se corresponden con los accidentes con víctimas notificados por los diferentes cuerpos de policía y guardia civil y 202.804 registros referentes a datos específicos de las propias víctimas. Mediante la combinación adecuada de esta información se crea un único conjunto de datos cuyos registros representan a los accidentados y contiene la información tanto de éstos como de las circunstancias del siniestro en el que estuvieron implicados.

4.3.1.1. Antecedentes: Metodología para el estudio de tablas de siniestralidad vial

Se presenta, en primer lugar, el estudio antecedente a la consideración de un marco bipolar en la clasificación de estos datos. El objetivo principal de este trabajo es la creación de una metodología para el estudio de una base de datos de accidentalidad vial a partir de un procedimiento semiautomático para facilitar el tratamiento de las tablas de datos de esta naturaleza que, con periodicidad anual, son recogidas por la Dirección General de Tráfico. Esta metodología se puede dividir en las etapas secuenciales siguientes:

- Preprocesamiento de los datos
  - Proponer un procedimiento para el estudio y recodificación de las variables que históricamente se consideran de influencia en la gravedad de las lesiones producidas.

    - Determinar las subpoblaciones objetivo mediante estudio de la posible segmentación por tipo de vehículo (bicicletas, motos, camiones, turismos, autobuses...) o por tipo de víctima (peatón, conductor, pasajero...).

- Determinación de factores de riesgo

    - Identificar y evaluar los factores de riesgo influyentes en los accidentes con víctimas mortales en las distintas subpoblaciones.
    - Determinar perfiles de víctimas y escenarios de accidentalidad como resultado de la integración de distintas técnicas de clasificación.
    - Abordar el problema de clasificación de las clases poco representadas. La proporción del evento considerado, la muerte a 30 días del siniestro, es únicamente del 1,2

- Modelos de clasificación en minería de datos

    - Crear funciones para facilitar el ajuste de algoritmos de aprendizaje a los datos para la clasificación supervisada del evento comentado y valoración de los resultados.
    - Establecer una comparativa de bondad de ajuste entre distintos métodos de clasificación en minería de datos tales como Random Forest, Gradient Boosting y Neural Networks, en la clasificación de los fallecidos en las distintas subpoblaciones de víctimas.

En definitiva, se pretende establecer un procedimiento mediante el cual sea posible extraer conclusiones de los datos de forma semiautomática y con la máxima fiabilidad posible. Hay que observar en este punto que se han utilizado únicamente los datos del año 2012, por lo que quedaría abierta la línea de investigación para el tratamiento de datos de naturaleza longitudinal. En cualquier caso, se considera que esta metodología proporciona una base sólida para obtener buenos resultados en los conjuntos de datos de accidentalidad vial recogidos de esta forma.

Para alcanzar los objetivos planteados se recurre a una metodología que incluye muchos aspectos del tratamiento de datos que se detallan a continuación.

Preprocesamiento. Para un correcto análisis del problema presentado, la fase previa a un análisis estadístico es una etapa de depuración de los datos, cuyo objetivo es minimizar el *ruido* que ciertas distribuciones de variables pueden introducir en los modelos, con la consecuente pérdida de precisión e incluso la posible obtención de conclusiones erróneas. Debido a esto, se realiza una cuidadosa labor de examen y depuración de las variables que resultarán

	Camiones	Bicis	Motos	Ciclom.	Peatones	Turismos
muert30 (% Si)	1.23	1.28	1.41	0.74	3.1	0.64
accseg (%No)	7.92	32.94	6.54	5.37	39.26	8.02
alcohol (% Si)	2	1.05	1.98	2.67	2.59	5.31
infracc (% Si)	35.97	46.03	47.78	50.46	41.08	38.59
distracc (% Si)	46.93	33.15	28.47	30.11	37.27	42.83
velina (% Si)	10.48	5.58	8.84	5.2	13.53	9.99
Num. Registros	5512	5535	20716	8535	11504	134878

Tabla 4.12: Distribución de las variables explicativas de mayor interés en las subpoblaciones.

de interés a lo largo del estudio. En cuanto a los métodos de recodificación de variables, por un lado y ya que el objetivo fundamental es preparar el conjunto de datos y sus variables para un análisis de clasificación binaria, se han llevado a cabo uniones de niveles de variables de naturaleza categórica que presentaban excesivo número de niveles o categorías irrelevantes. Para ello, se ajustan modelos de regresión logística univariante en los que la variable respuesta es el suceso de interés de este estudio, la variable muerte a 30 días de naturaleza dicotómica, y la explicativa es la variable a recodificar. De esta forma, se unifican categorías no significativas por sí mismas y con parámetros estimados de igual signo, ya que su influencia en el evento se da en el mismo sentido. El otro procedimiento de recategorización de variables utilizado han sido los árboles de clasificación CHAID (CHi-squared Automatic Interaction Detection).

En la Tabla 4.12 se presentan las distribuciones de las variables propias del conductor a modo de comparativa entre las distintas subpoblaciones para ilustrar la diferencia de distribuciones que justifica el estudio por separado. Se observan las variables (número de niveles), *Muerte a 30 días* (2), *Accesorios de seguridad* (2), *Alcohol* (2), *Infracción* (2), *Distracción* (2) y *Velocidad Inadecuada* (2) A las variables comentadas se une la Edad (4) que se recategoriza en cuatro tramos. Se pone de manifiesto la bajísima incidencia del evento de muerte a 30 días en la población, lo que augura una etapa de clasificación con complicaciones aseguradas.

En lo referente a los factores propios de la vía se han considerado los elementos presentes en ésta que pueden resultar peligrosos en los siniestros de vehículos de dos ruedas como *Mediana entre calzadas* (2), *Barrera de seguridad* (2), *Paneles direccionales* (2), *Hitos de arista* (2), *Captafaros* (2) o estado de la *Superficie* (4), teniendo también en cuenta variables como *Tipo* (3) y *Titularidad* (4) *de la vía*, *Densidad de circulación* (4) y *Zona* (4). También se consideran variables como *Factores Atmosféricos* (5), *Tipo de Accidente* (10), *Luminosidad* (4) o *Tipo de Día* (4).

Técnicas de clasificación empleadas. En lo que se refiere a modelos de clasificación, se presta especial atención a la Regresión Logística como la técnica clásica debido a su base estadística y a la posibilidad de cuantificar el efecto de las variables sobre la respuesta mediante los *odds ratio* (razón de probabilidades que cuantifica el número de veces que el riesgo de cierto evento sucede en una subpoblación o categoría de una variable), frente a la metodología utilizada por otros algoritmos de aprendizaje estadístico y computacional. A continuación se enumeran las distintas técnicas utilizadas:

La Regresión Logística [43] cuya gran ventaja es la posibilidad de cuantificar los efectos de los predictores sobre la respuesta a través de los *odds ratio*. Como algoritmos de machine learning se han utilizado: Redes Neuronales (ANN) [64], Random Forest (RF) [9], Gradient Boosting (GB) [30], Extreme Gradient Boosting (XGB) [13], Boosted Logistic Regression (LogiBoost) [29] y finalmente los Bayesian Generalized Linear Models (BayesGLM) [35], todos ellos disponibles en el paquete Caret [48] de R.

En todas las técnicas de minería de datos se ha aplicado el método de validación cruzada repetida. Particularmente, se aplica el esquema de validación cruzada *k-folds* definido en la Sección 4.2.1 al archivo en cada iteración del algoritmo. Este proceso de particiones se repite m veces consiguiendo por lo tanto $m \times k$ modelos ajustados con conjuntos de entrenamiento distintos y validado sobre observaciones no utilizadas en su construcción. Se ha elegido $k = 3$ y $m = 4$ en este estudio.

Ensamblado de modelos. Con el objetivo de mejorar la precisión alcanzada por los modelos de clasificación empleados en el estudio y reducir la varianza de los errores cometidos, se proponen distintos métodos de ensamblado de clasificadores mediante la técnica de *stacking*. Este método consiste en construir clasificadores dados por la combinación, lineal o no, de las probabilidades estimadas por los modelos ajustados, algunos de los cuales son ensembles en sí mismos (Random Forest, Gradient Boosting). Con ello se consiguen las probabilidades estimadas conjuntas, y se realiza la clasificación mediante la técnica del punto de corte óptimo de la probabilidad estimada. Cabe destacar que para obtener mejores resultados es conveniente realizar un estudio de correlaciones entre las predicciones para descartar los ensambles de probabilidades estimadas altamente correladas que usualmente no proporcionan mejora respecto al mejor modelo [22].

Factores de influencia en la mortalidad. Como resumen de este epígrafe, y tras el proceso de análisis de la importancia de las variables en los distintos modelos ajustados, se propone una medida de influencia de los distintos factores sobre el resultado fatal en accidentes de tráfico en las subpoblaciones de víctimas de siniestros viales en España en el año 2012. Se construye, para cada subpoblación del estudio, una tabla que contiene las

cinco variables más relevantes en cada uno de los modelos ajustados y se realiza un conteo de las frecuencias relativas de aparición de cada variable en el top 5 de la medida de importancia a lo largo de los ocho modelos.

Perfil	Escenario
Camiones	Edad$(38, 47]$-Acc.Seg-Distracc
Noche	Vuelco-Sal. Izq-Barrera-Mediana-Hitos
Bicis	Edad(>57)-Distracc-Infracc-Acc.Seg Festivo-Dens.Cir-Barrera
Motos	Edad(38,47]-Vel.Inadec-Infracc Colis.Front-Sal.Izq-Dens.Cir-Barrera
Ciclomotores	Edad(>38)-Acc.seg-Infracc-Distracc Festivo-Dens.Cir-Zona Urbana
Peatones	Edad(>57)-Distracc-Acc.Seg Noche- Festivo-Dens.Cir-Zona Urb.-Hitos-Barrera
Turismos	Edad(>57)-Distracc-Acc.Seg Dens.Cir-Colis.Frontal-Hitos-Barrera

Tabla 4.13: Perfiles de víctimas y escenarios de accidentalidad por subpoblaciones.

La idea fundamental es que las variables que aparecen con mayor frecuencia como importantes en los distintos modelos han de ser los factores que mayor influencia tienen sobre el suceso de interés, al haber sido seleccionados por distintos algoritmos para crear los modelos de clasificación.

En la Tabla 4.13 se presentan los perfiles de víctimas y escenarios de accidentalidad extraídos mediante este procedimiento para cada una de las subpoblaciones de interés. Se han puesto de manifiesto, mediante esta metodología, los factores tanto propios de la vía como inherentes al conductor que presentan una mayor influencia en la clasificación del evento de interés en el estudio. Cabe destacar las ventajas e inconvenientes de este método para la selección de factores de influencia, ya que por un lado supone un método robusto para esta tarea debido a que es un compendio de técnicas, y no un solo algoritmo, el que ha decidido extraer esa variables como influyente, evitando así posibles fallos o sesgos en la selección de variables de cada uno de ellos de manera individual.

Por otra parte, la principal desventaja de este método es la imposibilidad de cuantificar la importancia de estas variables así como el sentido de influencia en la clasificación del evento de interés debido a que se seleccionan para realizar la partición digamos en un nodo (en el caso de los métodos basados en árboles), pero es difícil saber si esa categoría de esa variable desemboca en un aumento o en un detrimento de la probabilidad estimada. Este hecho no tiene especial importancia ya que se dispone de métodos complementarios como la inspección descriptiva de la población y la interpretación de los OR de la regresión logística, con los que se puede decidir ese sentido de influencia.

Capacidad de clasificación. En este caso, debido a la baja incidencia del suceso de interés en la población, las probabilidades estimadas suelen

ser bajas y esto hace que el punto de corte que maximiza la relación entre sensibilidad y especificidad de la clasificación no sea el 0.5. Por ello, y para una mejor clasificación, se programan funciones de predicción que generan las matrices de confusión y se recurre a la función ROC para dibujar la curva y estimar el punto de corte óptimo para la probabilidad, aquel que hace máxima la suma de sensibilidad y especificidad para la probabilidad estimada frente a la clase real. Se comprueba que existe poca diferencia entre este valor estimado y la prevalencia a priori del evento aunque, pequeñas variaciones en el valor considerado producen grandes cambios en el número de mal clasificados. Este hecho es habitual en conjuntos de datos no balanceados debido a las bajas probabilidades estimadas que crean una frontera de decisión muy dispersa entre las clases a predecir, y esta poca consistencia de la estimación del punto de corte es el mayor inconveniente de la utilización de este método. En cualquier caso se considera muy superior la capacidad de clasificación del evento por medio de este procedimiento, máxime al tratarse de un suceso fatal que ha de ser evitado. Es lógico actuar bajo la premisa de que las consecuencias de la mala clasificación de los registros de la clase de interés resultan de mucha mayor gravedad, y por ello ha de relajarse el umbral para la especificidad incurriendo en una mayor tasa de falsos positivos. Realizando este procedimiento para todos los modelos en todas las subpoblaciones se obtienen los valores del estadístico AUC o área bajo la curva ROC y el punto de corte óptimo de la probabilidad estimada como aquel que maximiza la suma entre sensibilidad y especificidad. En total se han realizado ocho predicciones para cada una de las seis subpoblaciones, con lo que se han ajustado cuarenta y ocho modelos finales en estos datos escogidos de entre cientos probados.

	Log.	L_{Boost}	ANN	RF_{100}	RF_{500}	GBM	XGB	Bayes	Media
Camiones	.84	.79	.85	.7	.76	.85	.91	.86	.82
Motos	.88	.83	.93	.74	.78	.87	.9	.88	.85
Bicis	.9	.84	.94	.74	.82	.92	.93	.91	.88
Ciclos	.86	.78	.91	.76	.79	.87	.89	.86	.84
Peatones	.88	.85	.96	.66	.75	.93	.95	.9	.86
Turismos	.89	.86	.94	.72	.77	.93	.95	.9	.87
Media	.88	.83	.92	.72	.78	.90	.92	.89	

Tabla 4.14: Comparativa de precisión (ROC) global. Punto de corte óptimo

A la vista de los resultados de la Tabla 4.14, ANN es el algoritmo que mejor ajusta la clasificación en las subpoblaciones de Motos, Bicis, Peatones y Ciclomotores mientras que XGB lo hace en Camiones con mucha diferencia, y en Turismos con no tanta. En general se considera que el ajuste de los algoritmos de clasificación utilizados es muy bueno, alcanzando valores de AUC superiores al 0.9. Sin embargo, al tratarse de una población con clases no balanceadas es importante evaluar la sensibilidad y especificidad de la

clasificación.

	ROC	Sens.	Espec.
Camiones	.91	.91	.78
Motos	.93	.87	.87
Bicis	.94	.93	.83
Ciclomotores	.91	.95	.84
Peatones	.96	.85	.81
Turismos	.95	.93	.85
Media	.93	.91	.83

Tabla 4.15: Medidas de ajuste para el mejor modelo de cada subpoblación.

La información de la Tabla 4.15 refleja la elevada capacidad de los modelos ajustados para clasificar a los verdaderos eventos, con una sensibilidad que supera el 85 % en todos los casos, llegando, en el mejor de ellos (subpoblación de ciclomotores) al 95 %.

La elevada capacidad de clasificación de los modelos propuestos justifica la robustez del método de extracción de factores de riesgo comentado en el anterior epígrafe.

Resultados del ensamblado de modelos Como último apartado del estudio, se proponen aquí algunos métodos de ensamblado de los modelos anteriormente ajustados con el fin de obtener clasificadores cuya relación entre sensibilidad y especificidad sea mayor que la proporcionada por los modelos individuales. Para ello se construye un conjunto de datos que contiene las probabilidades estimadas por los ocho modelos ajustados para cada subpoblación con el fin de combinar estas probabilidades de distintas formas. Sin ánimo de profundizar en los modelos de ensamblado óptimos, se proponen varios de estos posibles clasificadores combinados y se compara su capacidad.

En primer lugar se construye un clasificador dado por la media aritmética de las probabilidades estimadas por cada uno de los modelos individuales, llamado ensamble medio (EnsMean). A continuación se realizará un ajuste de regresión logística por pasos, *backward*, con las ocho probabilidades estimadas como predictores para la clasificación del evento, y se construirá un clasificador que viene dado por la media ponderada por pesos obtenidos por los coeficientes de la regresión ajustados, de forma relativa. Este clasificador se llamará ensamble regresión (EnsRegw). Así mismo se considerará la probabilidad estimada de este modelo de regresión logística como otro posible ensamble logístico (EnsRegPred).

Por último, se construye un clásico ensamble dado por la media ponderada de Gradient Boosting y Random Forest, que se considera interesante debido a las distintas formas de actuación de ambos algoritmos, estando el

primero orientado a reducir el sesgo de las estimaciones y con la ventaja de la selección por sorteo de variables del segundo. Se consideran, tras diversas pruebas, los pesos de 0.8 y 0.2 respectivamente (Ens2080). Antes de crear los ensambles se realiza un estudio de correlaciones entre las probabilidades estimadas ya que conviene integrar los resultados de los modelos que presenten mayor independencia para contrarrestar errores de clasificación.

	EnsRegPred	EnsRegw	EnsMean	Ens2080
Camiones	.85	.89	.87	.85
Motos	.93	.91	.9	.89
Bicis	.92	.93	.92	.92
Ciclomotores	.96	.94	.93	.93
Peatones	.9	.89	.88	.87
Turismos	.93	.95	.94	.93

Tabla 4.16: Comparativa de precisión (ROC) para los modelos de ensamble.

Se puede observar en la Tabla 4.16 que el ensamblado de modelos presenta buenos ajustes a los datos de estudio, en especial en la subpoblación de ciclomotores, en la que se consigue un valor del área bajo la curva ROC de 0.96, siendo de 0.91 el valor del modelo base ganador en la comparativa del epígrafe anterior.

Conclusiones. La metodología de estudio propuesta responde correctamente a los objetivos planteados en cuanto al estudio descriptivo de la población y el preprocesamiento de los datos disponibles, así como a los dos grandes puntos de interés del estudio. La determinación de los factores, ya sean propios del conductor, de la vía o aquellos circunstanciales, que elevan la probabilidad de resultar fallecido en accidente de tráfico en esta población, con resultados que confirman lo ya obtenido en otros muchos estudios sobre accidentalidad. Por otra parte la comparativa de bondad de ajuste de modelos de machine learning para la clasificación supervisada de la variable de interés *muerte a 30 días*, que arroja buenos resultados siendo eXtreme Gradient Boosting (XGB) y Redes Neuronales (ANN) los algoritmos con mayor capacidad de clasificación.

4.3.1.2. Aplicación del paradigma bipolar en la clasificación multiclase de datos de siniestralidad vial

En el marco de problemas de clasificación con conjuntos de datos con clases no balanceadas y sus casos extremos se evalúa la mejora de resultados con la incorporación de información bipolar, siguiendo las ideas presentadas en el marco teórico de la Sección 4.1 para llevar a cabo el proceso de construcción del modelo de clasificación. En particular se aborda el caso de clasificación

mediante arboles con metodología CART. La agregación de información bipolar a los árboles de decisión para la clasificación supervisada multiclase comienza con la adición de la filosofía bipolar "a posteriori" en la clasificación. Tomando las probabilidades estimadas de pertenencia a cada una de las clases de interés, se evalúa la información de carácter positivo y negativo asociada a cada una de ellas y se agregan mediante un operador para obtener una nueva clasificación que recoja de forma más fiel los patrones subyacentes en los datos.

La aplicación de este procedimiento se realiza para la tarea de clasificación supervisada de la variable *lesividad* (categórica ordinal con 4 niveles) que recoge la gravedad de las lesiones producidas valoradas en el momento del siniestro con cuatro niveles, Ileso (I), Herido Leve (HL), Herido Grave (HG) y Muerto (M) y cuyas clases están altamente desbalanceadas. En la Tabla 4.17 se presentan los resultados preliminares de la aplicación a los datos con dos matrices de disimilitud, D1, considerada ad-hoc para el caso particular, y D2, la propuesta en [69, 70].

Tipo	Accuracy	Kappa	TVP	Data	Disim.
Arbol	.4464	.2618	.5405	Balan	D1
ArbolBip	.4453	.2605	**.5445**	Balan	D1
Arbol	.4464	.2618	.5405	Balan	D2
ArbolBip	.4317	.2422	**.7267**	Balan	D2
Arbol	.5787	.1874	.0455	NoBalan	D1
ArbolBip	.5792	**.1914**	**.0682**	NoBalan	D1
Arbol	.5787	.1874	.0455	NoBalan	D2
ArbolBip	.5751	.1784	**.1212**	NoBalan	D2

Tabla 4.17: Comparativa de precisión de árboles y árboles con información bipolar.

Por tanto, la metodología descrita se aplica al conjunto de datos balanceado y no balanceado, y se computan medidas clásicas de ajuste (*accuracy* y *kappa*) , y la Tasa de Verdaderos Positivos (TVP) considerada como la suma de Muertos y Heridos Graves. Estas medidas se describen con detalle en la Sección 2.7.

En la Tabla 4.17 se pueden apreciar ligeras mejoras especialmente cuando se considera el conjunto de datos original sin recurrir al re-balanceo de los datos. En todos los casos ambas estructuras de disimilitud consiguen aumentar la capacidad de reconocimiento de instancias pertenecientes a la clase positiva por lo que mejora la TVP.

Conclusiones. La incorporación de un marco de representación bipolar de conocimiento a la clasificación de la variable categórica de carácter ordinal *lesividad*, que determina la gravedad del accidente, permite un tratamiento

más flexible del problema de clasificación de datos de siniestralidad observándose, en esta etapa inicial, la ligera mejora que aporta su consideración a posteriori respecto al árbol de clasificación base. Cabe destacar que el marco aplicado aquí, considera estructuras de disimilitud pre-fijadas, no aplicándose el paradigma de aprendizaje propuesto en la Sección 3.5. Por ello, los resultados obtenidos son susceptibles de mejora concediendo libertad al clasificador para aprender la estructura disímil de los propios datos. Esta línea futura se comentará detalladamente en la Sección 8.2.

4.3.2. Clasificación supervisada para la detección de bordes en imágenes

Tradicionalmente, el último paso de los algoritmos de detección de bordes, que se denomina escalado-evaluación, produce la salida final clasificando cada píxel como *borde* o *no borde*. Este último paso generalmente se realiza en base a métodos de evaluación locales. La evaluación local realiza esta clasificación basándose en las medidas obtenidas para cada píxel. Por el contrario, en este trabajo, se considera un enfoque de evaluación global basado en la idea de una lista de bordes para producir una solución que se adapte más a la percepción humana. En particular, se propone un nuevo método de evaluación que se puede combinar con cualquier algoritmo de detección de bordes clásico de una manera fácil para producir un borde novedoso algoritmo de detección y clasificación de bordes.

El nuevo método de evaluación global se divide en cuatro pasos: en primer lugar, se construyen la lista de bordes, denominados segmentos de borde. En segundo lugar, se extraen las características asociadas a cada segmento: longitud, intensidad, ubicación, etc. En el tercer paso, se aprenden las características que hacen que un segmento sea lo suficientemente bueno para convertirse en *borde*. En el cuarto paso, se aplica lo anterior a la tarea de clasificación.

Es especialmente destacable en este trabajo la creación de la variable de supervisión o *ground truth* asociada a la lista de bordes, que resulta necesaria para la clasificación supervisada. Dado que las tareas asociadas al análisis de imágenes están del todo supeditadas a la percepción humana, no es sencillo propósito la representación exacta de la *verdad*, si es que tiene sentido hablar de ella en estos términos. En efecto, en el contexto concreto de la detección de bordes en imágenes, resulta compleja tarea encontrar coincidencias significativas en los juicios humanos a este respecto. En primer lugar, distintas personas realizan interpretaciones diferentes del significado de la palabra *borde* y, mas allá, perciben la imagen y la tarea asociada a ella de formas divergentes. Así, los juicios de distintos humanos para una misma imagen pueden diferir sustancialmente. Es por ello que este tarea se enmarca en el campo de la decisión multi-criterio, donde realmente existe más de una variable de supervisión.

Con el fin de conseguir la clasificación final nítida, se realiza una agregación de los juicios de los humanos consultados que actúa como variable objetivo en un posterior problema de clasificación con clases poco representadas, abordado mediante el ajuste de los clasificadores probabilísticos presentados en la Sección 2.5 bajo distintos paradigmas a nivel de muestra y a nivel de algoritmos. Finalmente, se prueba la efectividad de este algoritmo contra otros algoritmos clásicos basados en un enfoque de evaluación local, obteniendo mejoras significativas a nivel estadístico.

Los pormenores de este trabajo trascienden en gran parte los objetivos planteados en esta memoria. Información adicional sobre el marco teórico y los resultados obtenidos puede ser consultado en [26, 27].

Capítulo 5

Representación Bipolar del Conocimiento para el desarrollo de nuevos algoritmos de clasificación difusos

No hay nada peor que la imagen nítida
de un concepto difuso.

Ansel Adams

Resumen: Se reserva este capítulo a la presentación de los avances obtenidos de la aplicación del paradigma de RBC en el contexto de clasificadores de tipo difuso. Se hace necesario, en primer lugar, una particularización (ver Sección 5.1) del marco general bipolar presentado en el Capítulo 3 para este caso concreto en el que la evidencia viene dada, de una u otra forma, por grados de pertenencia difusos. En la Sección 5.2 se detallan los resultados alcanzados mediante la aplicación de un paradigma bipolar a nivel global sobre clasificadores robustos construidos a partir de un esquema replicación-agregación de algoritmos probabilísticos.

5.1. Marco teórico bipolar-difuso

En este capítulo se explora y evalúa el comportamiento de los métodos de RBC presentados en este trabajo, en el contexto de los clasificadores difusos definidos en la Sección 2.4 del capítulo de conceptos preliminares.

Recordemos que, de acuerdo a la Ecuación (2.3), todo clasificador difuso C_F puede ser visto como una función que asigna a cada instancia x y cada clase C_i con $i \in \{1 \dots c\}$, un valor numérico $\mu_i \in [0,1]$ que representa el grado de pertenencia del objeto a dicha clase. Así, un clasificador difuso, en la etapa previa a la asignación de clase, alcanza para cada x un vector de grados de pertenencia $C_F(x) = (\mu_1, \dots, \mu_c)$ que, a diferencia del caso probabilístico detallado en la Sección 4.1, no se encuentran sujetos a la restricción $\sum_{i=1}^{c} \mu_i = 1$.

Es entonces cuando, considerando el vector de grados de pertenencia $C_F(x) = (\mu_1, \dots, \mu_c)$ obtenido por el clasificador como el vector de evidencia $ev(x) = (ev_1(x), \dots, ev_c(x))$ presentado en la Sección 3.1 del Capítulo 3, cualquier clasificador difuso es susceptible de ser enriquecido mediante la aplicación de un marco de RBC.

Por tanto, es relevante presentar la particularización del marco general de aplicación de la RBC en el contexto de clasificadores de naturaleza *soft* descrito a lo largo del Capítulo 3, en el marco específico de los algoritmos de clasificación supervisada de tipo difuso. Así, en adelante se considera que el vector de grados de pertenencia $C_F(x) = (\mu_1, \dots, \mu_c)$ dado por el algoritmo, representa la evidencia de carácter positivo en el sentido propuesto en la Sección 3.1, por lo que denotamos $\mu^+ = \left(\mu_1^+, \dots, \mu_c^+\right) = (\mu_1, \dots, \mu_c)$, que en este caso no necesariamente satisface la restricción $\sum_{i=1}^{c} \mu_i^+ = 1$.

A continuación, se presenta la extensión de los procedimientos de generación (Sección 5.1.1) y agregación (Sección 5.1.2) de evidencia bipolar detallados en las Secciones 3.2 y 3.3.1, respectivamente.

5.1.1. Obtención de los pares (μ^+, μ^-)

Es en este punto donde, siguiendo el esquema presentado en la Sección 3.2, se representa por μ_i^- el grado de de pertenencia de una instancia x a la clase disímil a C_i, entendida como ese constructo abstracto representando el conjunto de clases consideradas disímiles a C_i. Así mismo se denota por $D = (d_{ij})$ la matriz que cuantifica la estructura de disimilitud en el conjunto de clases, donde d_{ij} se entiende como el grado de disimilitud entre las clases C_i y C_j. Obviamente, se consideran matrices de disimilitud en la forma presentada en la Sección 3.1, es decir satisfaciendo D, $d_{ii} = 0$ para todo $i = 1, \dots, c$ y siendo en general no simétricas.

Tomando en cuenta la estructura de disimilitud y el vector de grados de pertenencia considerados como evidencia positiva, se construye la evidencia de carácter negativo μ_i^- como,

$$\mu_i^-(x) = \sum_{j \neq i} d_{ij}\ \mu_j^+(x) = \sum_{j=1}^{c} d_{ij}\ \mu_j^+(x) = D_i \mu^+(x) \tag{5.1}$$

o en forma matricial,

$$\mu^-(x) = D\mu^+(x) \tag{5.2}$$

Observación 4. Nótese que, en el caso en que, para una cierta clase C_i, el grado de disimilitud entre C_i y el resto de clases en el conjunto es máximo, $d_{ir} = 1$ para todo $r \neq i$, entonces $\mu_i^- = \sum_{j=1}^{c} \mu_j^+ - \mu_i^+$ (ya que no ha de satisfacerse la condición de suma unidad, $\sum_{j=1}^{c} \mu_j^+ \neq 1$) que es precisamente la representación del caso clásico en el que el grado de pertenencia negativo es el grado de no pertenencia a la clase C_i. En general y, con mayor frecuencia en estas situaciones donde existe una confusión entre las clases ($d_{ir} < 1$), se tiene que $\mu_i^- < \sum_{j=1}^{c} \mu_j^+ - \mu_i^+$.

En este sentido, se presenta el proceso de obtención de los pares de información bipolar en el contexto de los clasificadores de naturaleza difusa, con la siguiente estructura:

$$\begin{array}{rccccc} C_F^{bip} : X & \longrightarrow & [0,1]^c & \longrightarrow & [0,1]^c \times [0,1]^c \\ x & \longrightarrow & C_F(x) = (\mu_1, \ldots, \mu_c) & \longrightarrow & (\mu_1^+, \ldots, \mu_c^+) \times (\mu_1^-, \ldots, \mu_c^-) \end{array}$$

Figura 5.1: Esquema de construcción de información bipolar. Primera etapa (E1) de un *Clasificador Bipolar Difuso Ajustado* C_F^{bip}

En la Figura 5.1, C^{bip} representa un clasificador bipolar difuso, en cuya primera etapa (E1), partiendo de un item $x \subset X$ extrae el vector de grados de pertenencia $(\mu_1, \ldots, \mu_c) \in [0,1]^c$ y genera los pares de evidencias (en este caso grados de pertenencia) de carácter positivo y negativo $(\mu_1^+, \ldots, \mu_c^+) \times (\mu_1^-, \ldots, \mu_c^-) \in [0,1]^c \times [0,1]^c$.

Toda vez la estructura bipolar pareada ha sido obtenida, uno de entre las muchos métodos de explotación posibles es la utilización de operadores de agregación. En este sentido, se presenta en la Sección 5.1.2, la particularización del marco general de agregación introducido en la Sección 3.3.1 al caso de clasificadores difusos.

5.1.2. Agregación de los pares (μ^+, μ^-)

Como se ha señalado en la Sección 3.3.1, en el campo de los operadores de agregación, existe gran variedad de funciones capaces de representar la información contenida en los pares de grados de pertenencia bipolares (μ_i^+, μ_i^-), $i \in \{C_1, \ldots, C_c\}$, mediante un solo valor numérico, que se puede denotar en este caso como μ_i^{adj}, representando la grado de pertenencia agregado o ajustado. Por tanto, ahora la cuestión clave es la elección de operadores de agregación que unifiquen de forma adecuada las informaciones de carácter positivo y negativo.

$$\begin{array}{ccccc} [0,1]^c \times [0,1]^c & \longrightarrow & [0,1]^c & \longrightarrow & \{C_1,\ldots,C_c\} \\ (\mu_1^+,\ldots,\mu_c^+) \times (\mu_1^-,\ldots,\mu_c^-) & \longrightarrow & (\mu_1^{adj},\ldots,\mu_c^{adj}) & \longrightarrow & C_{arg\,\max\{\mu_1^{adj},\ldots,\mu_c^{adj}\}} \end{array}$$

Figura 5.2: Esquema de agregación bipolar. Segunda etapa (E2) de un *Clasificador Bipolar Difuso Ajustado* C_F^{bip}

En la Figura 5.2, se muestra el diagrama de flujo de la segunda etapa (E2) de lo que se ha llamado *Clasificador Bipolar Difuso Ajustado* C_F^{bip}. Este proceso comienza con la consideración de los anteriormente calculados pares de grados de pertenencia $(\mu_1^+,\ldots,\mu_c^+) \times (\mu_1^-,\ldots,\mu_c^-) \in [0,1]^c \times [0,1]^c$ que, mediante la aplicación de un operador de agregación quedan representados por un único vector de grados de pertenencia agregados $(\mu_1^{adj},\ldots,\mu_c^{adj})$. La decisión nítida final es llevada a cabo mediante la conocida regla del máximo, seleccionándose la clase C_i con mayor valor de grado de pertenencia agregado.

En lo que sigue se definen los dos operadores de agregación ya estudiados en la Sección 3.3.1, ahora en un marco difuso. En primer lugar, el operador *Aditivo* difuso se define como sigue.

Definición 5.1.1. Sean μ_i^+, μ_i^- los grados de pertenencia positivos y negativos de una instancia a la clase C_i. Se defines el Grado de Pertenencia Aditivo del objeto x a la clase C_i como

$$\mu_i^{add} = \max\{0, \mu_i^+ - \mu_i^-\}$$

Es preciso interpretar la información dada por el operador aditivo previamente definido, de modo que un valor $\mu_i^{add} > 0$ representa un salto o discrepancia entre la grado de pertenencia a la clase C_i y la grado de pertenencia al constructo definido como *clase disímil* a C_i. Por el contrario, un valor nulo de μ_i^{add} es un indicador de la existencia de mayor cantidad de afectos negativos que positivos en la elección de dicha clase.

En la siguiente definición se presenta un método alternativo de agregación de las informaciones positiva y negativa en una sola puntuación que hemos llamado *grado de pertenencia ajustado logístico* de acuerdo a la notación usualmente utilizada para la función de agregación de tipo exponencial utilizada.

Definición 5.1.2. Sean μ_i^+, μ_i^- los grados de pertenencia positivos y negativos de una instancia a la clase C_i. Se define el Grado de Pertenencia Ajustado Logístico del objeto x a la clase C_i como

$$\mu_i^{log} = \begin{cases} 1 - e^{-\frac{\mu_i^+}{\mu_i^-}} & si\ \mu_i^- > 0 \\ 1 & en\ otro\ caso \end{cases}$$

Como se ha destacado en la Sección 3.3.1 la agregación logística está basada en el ratio entre informaciones positiva y negativa, ajustando éste al rango $[0, 1]$ a través de una transformación logística, por lo que este esquema permite un comportamiento de alguna forma más flexible de los grados ajustados que el proporcionado por la agregación aditiva.

Para finalizar el proceso llevado a cabo por el *Clasificador Bipolar Difuso Ajustado* CF^{bip}, tras la aplicación de alguna de las agregaciones anteriormente propuestas y una vez los grados de pertenencia bipolares ajustados $\mu_i^{adj}(x)$ han sido obtenidos para cada clase (ya sea $\mu_i^{adj}(x) = \mu_i^{add}(x)$ o $\mu_i^{adj}(x) = \mu_i^{log}(x)$), la decisión nítida final en la clasificación del ítem x se lleva a cabo por aplicación de la regla del máximo sobre estos grados bipolares ajustados como se observa en el esquema de la Figura 5.2.

Para concluir la sección, y como concepto general que engloba el proceso descrito, se define formalmente un *Clasificador Bipolar Difuso Ajustado* C_F^{bip} a continuación.

Definición 5.1.3. Dado un clasificador difuso $C_F : X \longrightarrow [0, 1]^c$ en un universo de discurso X y dada una matriz de disimilitud D que permita crear la información bipolar $(\mu^+, \mu^-) = \big((\mu_1^+, \mu_1^-), \ldots, (\mu_k^+, \mu_c^-)\big)$ para cada instancia x a partir de la información dada por un clasificador difuso $C_F(x) = (\mu_1, \ldots, \mu_c)$. El *Clasificador Bipolar Difuso Ajustado* se define como

$$C_F^{bip}(x) = C_h \text{ si y solo si } \mu_h^{adj} = \text{máx}\{\mu_j^{adj}; j = 1 \ldots c\}.$$

donde μ^{adj} es el grado de pertenencia ajustado resultado de la agregación de μ^+ y μ^-.

De forma general, se denotarán en adelante como C_F^{bipAdd} y C_F^{bipLog} los *Clasificadores Bipolares Difusos Ajustados* aditivo y logístico respectivamente, construidos a partir de la información dada por cierto clasificador difuso C_F.

Observación 5. Cabe destacar que, el marco general de aplicación de la RBC no se restringe al uso de operadores de agregación para la explotación de la evidencia (grados de pertenencia difusos en este caso) bipolar. Por el contrario, distintos tipos de esquemas de explotación podrían ser considerados. En este sentido, se extiende la definición anterior teniendo en cuenta que ahora $C_F^{bip}(x) = C_h$ tal que $\Phi(\mu^+, \mu^-) = C_h$, siendo Φ la función de explotación general considerada introducida en la Sección 3.3.2.

5.2. Aplicación sobre clasificadores difusos robustos construídos a partir de clasificadores probabilísticos

Debido al esquema de entrenamiento paramétrico de los clasificadores probabilísticos considerado en gran parte de este trabajo, en el que se aplica una replicación bootstrap para la elección de los parámetros óptimos en la muestra de entrenamiento (ver Secciones 4.2.2, 4.2.3), muchos son los factores no controlados (semillas, aleatoriedad, inicialización, ...) que pueden influir en el desarrollo y el rendimiento de un proceso de clasificación supervisado una vez fijado el conjunto de datos de entrenamiento . Por tanto, incluso para el mismo conjunto de datos de entrenamiento , la estimación de las diferentes probabilidades $C_P(x) = (C_P(x)_1, \dots C_P(x)_c)$ que un clasificador probabilístico proporciona para cada objeto $x \in X$ en el conjunto de datos de entrenamiento podría variar dependiendo de la realización particular del algoritmo en el paso de entrenamiento. Teniendo en cuenta esta consideración, cuando se aplica la *replicación* estadística para mejorar la robustez del proceso de clasificación de un clasificador probabilístico, es posible interpretar la imprecisión que surge en la estimación de probabilidad como un conjunto difuso.

5.2.1. Configuración experimental

La Figura 5.3 muestra un diagrama de flujo de la etapa de entrenamiento propuesta (S1). Notemos que en caso de que nuestro clasificador ya sea difuso, esta etapa no presenta cambios particulares con respecto a la fase de entrenamiento habitual.

Se proporciona aquí un procedimiento para lograr esta *fuzzificación* orientada a la mejora de la robustez de un clasificador probabilístico. Se asume que la estimación de las probabilidades $C_P(x) = (C_P(x)_1, \dots, C_P(x)_c)$ se replica de alguna manera un número determinado de veces (por ejemplo, utilización de diferentes semillas para obtener diferentes muestras *bootstrap* del conjunto de datos de entrenamiento , o permitir una inicialización diferente del algoritmo) para obtener una evaluación de la imprecisión de la estimación involucrada. Sea L el número de repeticiones aplicadas, y $C_P^l(x) = (C_P^l(x)_1, \dots, C_P^l(x)_c)$ la distribución de probabilidad para las distintas clases obtenidas en la realización l-ésima del clasificador probabilístico. Así, la idea es agregar las distintas distribuciones de probabilidad $C_P^l, l = 1, \dots, L$ alcanzadas tras la replicación para obtener los grados de pertenencia del objeto x a cada clase. En este punto se consideran tres operadores de agregación para construir el clasificador difuso agregado: el *máximo* (que podría considerarse como una medida de posibilidad en el sentido descrito en [19]), el *mínimo* y la *media aritmética* (que podría considerarse

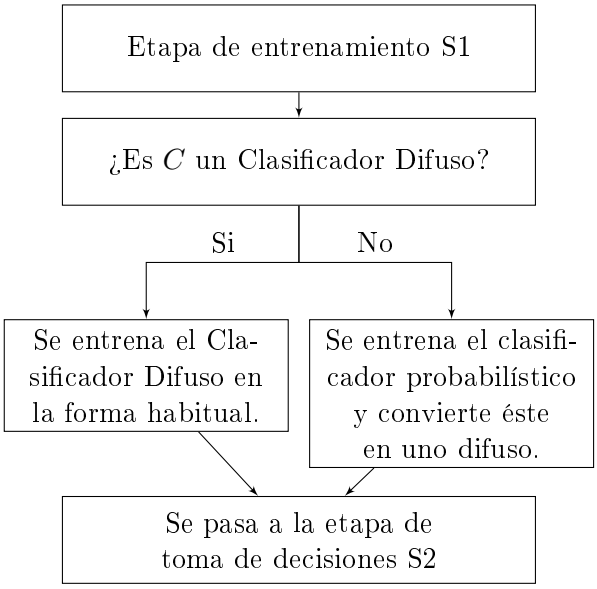


Figura 5.3: Diagrama de flujo de la Etapa de Entrenamiento propuesta (S1)

como una estimación más precisa de la probabilidad real de las diferentes clases).

Formalmente, digamos que ϕ denota cualquiera de estos tres operadores de agregación. Luego, los grados de pertenencia robustos del objeto x en la clase C_i se obtienen como

$$\mu_{C_i}(x) = \phi(C_P^1(x)_i, \ldots, C_P^L(x)_i). \tag{5.3}$$

Por lo tanto, se obtiene un clasificador difuso C_F en el sentido expuesto en la Ecuación (2.3) de la Sección 2.4 tras la agregación de las distintas distribuciones de probabilidad replicadas.

Los clasificadores base utilizados en este experimento son CART, Random Forest (RF) y Redes Neuronales (ANN). Este experimento está diseñado para comparar el rendimiento de estos clasificadores base, de los clasificadores difusos robustos obtenidos mediante la replicación y agregación de tales clasificadores base y de los clasificadores obtenidos a partir de los últimos mediante el proceso de aprendizaje de disimilitud propuesto bajo lo esquemas de agregación aditivo y logístico. Se calculan los diferentes resultados para cada uno de los tres operadores de agregación (máx, min y media) utilizados en los clasificadores base replicados. Por lo tanto, en realidad estamos llevando a cabo 9 experimentos diferentes (3 clasificadores base por 3 operadores de agregación), en cada uno de los cuales se comparan los rendimientos de 4 clasificadores (clasificador base, clasificador difuso fuzzificado, clasificador bipolar aditivo y clasificador bipolar logístico) utilizando varios conjuntos de datos. Por lo tanto, el objetivo es descubrir si los ajustes propuestos basados

en la disimilitud permiten una mejora del rendimiento de los clasificadores probabilístico base y difuso agregado robusto, bajo diferentes condiciones de agregación y para diferentes clasificadores base.

Los resultados para cada clasificador en cada experimento se obtendrán siguiendo un esquema de validación cruzada *5-folds* (ver Sección 4.2) para cada conjunto de datos. En cada partición *train/test*, es decir, para cada *fold*, la configuración paramétrica del clasificador base óptimo (parámetro de complejidad, c_p, para CART, número de variables seleccionadas al azar, *mtry*, en caso de RF y número de neuronas, *size*, o parámetro de disminución de peso, *decay*, para ANN) se aproxima usando una cuadrícula P en el espacio de parámetros de los algoritmos considerados. Para evaluar el rendimiento de cada configuración paramétrica específica $p \in P$, se generan 25 muestras *bootstrap* del conjunto de entrenamiento , de tal manera que los clasificadores básicos se ajustan a cada una de estas muestras y luego se prueban en un esquema OBB (*Out of Bag*, esto es, compuesta por las instancias no seleccionadas en cada muestra *bootstrap*), utilizando el estadístico kappa. El rendimiento de los clasificadores para cada configuración p (en cada *fold* de cada conjunto de datos) se computa promediando los valores de kappa obtenidos en cada muestra, y por lo tanto la configuración paramétrica p que se aplicará finalmente en el *fold* actual es aquella con la máxima media del valor kappa. Este clasificador de configuración paramétrica óptima se ajusta a todo el conjunto de entrenamiento y luego se aplica al conjunto de prueba, proporcionando vectores de probabilidades $C_P(x)$ para cada elemento x en los conjuntos de entrenamiento y prueba de cada *fold*. Cabe destacar que estos vectores pueden ser explotados a través de la regla del máximo para obtener asignaciones de clase nítidas, a partir de las cuales se pueden obtener los índices de rendimiento en los conjuntos de entrenamiento y prueba de los clasificadores probabilísticos básicos para el *fold* actual.

Replicamos este paso de entrenamiento de búsqueda paramétrica $L = 5$ veces en cada *fold* usando diferentes semillas para extraer las muestras de *bootstrap*, proporcionando así estimaciones de probabilidad replicadas $C_P^l(x)$, $l = 1, \ldots, L$ para cada objeto x en los conjuntos de entrenamiento y prueba. Luego aplicamos un operador de agregación ϕ (máximo, mínimo o media aritmética) a estas probabilidades replicadas, lo que proporciona los grados $\mu_{C_i}(x) = \phi(C_P^1(x)_i, \ldots, C_P^5(x)_i)$ del clasificador difuso agregado resultante C_F, evaluando la pertenencia de cada instancia de entrenamiento y prueba x para cada clase $i = 1, \ldots, c$. Como antes, estos grados difusos agregados pueden luego ser explotados a través de la regla del máximo para proporcionar asignaciones de clase nítidas y computar los índices de rendimiento de los conjuntos de entrenamiento y prueba de los clasificadores agregados difusos.

Las medidas de rendimiento en entrenamiento y prueba de cada uno de los 4 clasificadores en cada conjunto de datos considerado en cada ex-

perimento se calculan finalmente promediando los índices de kappa de los distintos *folds*. Para mayor claridad, los diferentes pasos de la configuración experimental descrita para cada conjunto de datos y clasificador de base probabilístico se resumen en el Algoritmo 1. Se puede observar que el procedimiento de ajuste del clasificador probabilístico a cada conjunto de datos se ha implementado utilizando el paquete *caret* (consulte [48]) del software R. En particular, las opciones predeterminadas de este paquete son $B = 25$ muestras *bootstrap* y $L = 5$ réplicas. Con respecto al espacio de los parámetros considerados, depende de las opciones disponibles para cada clasificador en las implementaciones específicas utilizadas para ajustar el modelo al conjunto de entrenamiento .

En lo concerniente a la determinación de la estructura disímil, en este trabajo se utiliza el esquema de aprendizaje evolutivo mostrado en la Sección 3.5, para obtener la estructura de disimilitud más conveniente del conjunto de clases para cada problema de clasificación.

Una característica importante de este novedoso enfoque es su independencia del marco de representación del conocimiento y la estructura de los clasificadores difusos, que permite su aplicación a cualquier algoritmo de esta clase, a pesar de la naturaleza y los pasos específicos intermedios que proporciona las puntuaciones difusas. En este sentido, el enfoque propuesto puede entenderse como un método general de post-procesamiento para ajustar la regla del máximo que generalmente se aplica para tomar la decisión sobre la asignación de clase asignada a cada elemento a clasificar.

Conjuntos de datos. Se seleccionan esta vez un banco de 25 conjuntos de datos del repositorio KEEL [75]. En particular, hemos utilizado los conjuntos de datos de validación cruzada de 5 *folds* proporcionados por KEEL en los diferentes experimentos. La Tabla 5.1 resume las propiedades de los conjuntos de datos seleccionados, mostrando para cada conjunto de datos el número de ejemplos (#Ex.), el número de atributos (#Atts.), su tipo (Real/Entero/Nominal) y la relación de desequilibrio (#IR) una vez que el conjunto de datos se ha transformado en un problema de clasificacion binaria. Para convertir un conjunto de datos de varias clases en un conjunto de datos desbalanceados de dos clases (C_1, C_2), hemos tomado como clase C_2 el original más cercano a 20 % de instancias y como clase C_1 la unión de las clases restantes.

Debemos señalar que muchos de los conjuntos de datos considerados presentan un alto IR, debido a su anterior distribución desequilibrada de múltiples clases. Se considera este hecho como una oportunidad para evaluar nuestros enfoques bipolares cuando se trata de conjuntos de datos desbalanceados.

Algoritmo 1 Procedimiento para la aplicación del esquema de RBC sobre clasificadores difusos robustos obtenidos como agregación de clasificadores probabilísticos

Seleccionar el número de *folds* F, réplicas L, y muestras *bootstrap* B y la rejilla P de configuraciones paramétrica para el entrenamiento del clasificador base probabilístico C_P.
for *cada fold* $f \in F$ **do**
  for *cada réplica* $l \in L$ **do**
    procedure Ajustar el clasificador base probabilístico al conjunto de entrenamiento (P, B)
      for *cada configuración paramétrica* $p \in P$ **do**
        for *cada muestra* $s \in S$ **do**
          Generar las muestras de prueba específicas
          Ajustar el clasificador a las restantes muestras
          Predecir la clasificación en las muestras de prueba
        end for
        Calcular el promedio de rendimiento a través de las B *muestras*
      end for
      Determinar la configuración paramétrica óptima
      Ajustar el clasificador al conjunto de entrenamiento completo considerando la configuración paramétrica óptima
      Obtener el vector de probabilidades estimadas $C_P^l(x)$
    end procedure
  end for
  Aplicar el operador de agregación ϕ *sobre los* L *vectores de probabilidad* $C_P^l(x)$
  Obtener los grados difusos agregrados $C_F(x)$
  for *cada agregación* $\in$ *{aditiva, logística}* **do**
    Recurrir al AG para aprender la estructura de disimilitud en el conjunto de entrenamiento
    Aplicar la matriz resultante D *sobre las puntuaciones* $C_F(x)$ *y obtener los grados bipolares* $C_{bip}(x)$
    Aplicar el método de agregación para obtener los correspondientes grados difusos ajustados $\mu^{adj}(x)$
  end for
end for
Calcular el promedio de rendimiento en los conjuntos de entrenamiento y prueba a lo largo de los F *folds para los* 4 *clasificadores una vez aplicada la regla del máximo sobre los grados difusos ajustados.*

Id.	Data-set	Clase C_2 (+)	#Ex.	#Atts.	(R/I/N)	#IR
App	Appendicitis	0	106	7	(7/0/0)	4.25
Aus	Australian	0	690	14	(3/5/6)	1.25
Bal	Balance	L	625	4	(4/0/10)	1.19
Bup	Bupa	1	345	6	(1/5/0)	1.38
Car	Car	acc	159	25	(15/0/10)	3.5
Con	Contraceptive	2	1473	9	(6/0/3)	3.43
Der	Dermathology	3	366	34	(0/34/0)	4.07
Eco	ecoli	im	336	7	(7/0/0)	3.34
Fla	flare	C	1066	25	(15/0/10)	2.58
Ger	German	2	1000	20	(0/7/13)	2.33
Gla	Glass	1	214	9	(9/0/0)	2.05
Hab	Haberman	positive	306	3	(0/3/0)	2.81
Hay	Hayes-Roth	3	160	4	(0/4/0)	3.76
Lin	Lymphography	malign_lymph	148	18	(3/0/15)	1.43
Nty	Newthyroid	2	215	5	(4/1/0)	5.14
Pag	Page-blocks	2	5472	10	(4/6/0)	14.57
Pen	Penbased	0	10992	16	(0/16/0)	8.57
Pim	Pima	tested_negative	768	8	(8/0/0)	1.88
Sah	Saheart	0	462	9	(5/3/1)	1.89
Shu	Shuttle	4	2175	9	(0/9/10)	5.44
Thy	Thyroid	2	720	21	(6/0/15)	18.1
Veh	Vehicle	positive	846	18	(0/18/0)	2.93
Wis	Wisconsin	2	699	9	(0/9/0)	1.83
Wnq	Winequality-red	7	1599	11	(11/0/0)	6.97
Yea	Yeast	MIT	1484	8	(8/0/0)	5.1

Tabla 5.1: Descripción de conjuntos de datos utilizados en la propuesta bipolar difusa robusta.

5.2.2. Resultados Experimentales

En adelante, se presentan los resultados del experimento computacional descrito anteriormente y llevado a cabo para estudiar la capacidad de mejora de nuestros clasificadores ajustados bipolares con respecto a los clasificadores base de referencia, así como a los clasificadores difusos obtenidos por el proceso de agregación, a los que se aplica el método de ajuste de decisión final propuesto.

Teniendo en cuenta que hemos considerado tres funciones de agregación diferentes, los resultados se mostrarán para cada una por separado. Por lo tanto, tendremos tres grupos de resultados según el método de agregación seguido para alcanzar la información difusa de los clasificadores probabilísticos replicados, y en cada grupo de resultados proporcionaremos una tabla principal de métricas de clasificación kappa acompañada de las observaciones de mayor relevancia.

Los resultados se agrupan, para cada algoritmo base, en pares para entrenamiento y prueba, donde el mejor resultado global para cada conjunto

de datos considerado se destaca en **negrita**. No se destaca ninguno en caso de empate.

5.2.2.1. Operador de agregación *Máximo*

Los resultados de las Tablas 5.2, 5.3 y 5.4 evidencian el buen comportamiento general del paradigma bipolar propuesto, al menos con respecto a uno de los métodos de ajuste bipolar, ya que permite mejorar el rendimiento tanto de los algoritmos de referencia como de los clasificadores agregados difusos.

	CART Max							
	Ref		Aggr		bipAdd		bipLog	
	Tr.	*Tst*	*Tr.*	*Tst*	*Tr.*	*Tst*	*Tr.*	*Tst*
Car	.796	.727	.795	**.733**	.797	.732	.797	.732
Con	.391	.247	.336	.232	.409	.289	.409	**.305**
Eco	.800	.642	.791	.595	.797	.642	.797	.642
Fla	.620	.560	.604	.559	.627	**.561**	.627	**.561**
Gla	.722	**.488**	.720	.473	.726	.482	.726	.482
Lin	.667	.620	.667	.620	.667	.620	.667	.620
Shu	.994	.995	.994	.995	.994	.995	.994	.995
Thy	.876	**.819**	.876	**.819**	.879	.806	.879	.806
Yea	.605	.474	.574	.480	.614	**.482**	.614	**.482**
Aus	.785	.673	.782	.684	.795	**.691**	.795	**.691**
Bal	.655	.518	.655	.518	.655	.518	.655	.518
Bup	.623	**.346**	.618	.344	.622	.345	.622	.345
Ger	.503	.301	.518	.335	.538	**.341**	.540	.338
Hay	.841	.738	.841	.738	.841	.738	.841	.738
Pen	.970	.932	.970	.932	.970	.932	.972	**.939**
Pag	.872	.887	.872	.887	.872	.887	.872	.887
Sah	.545	.220	.511	**.340**	.582	.266	.582	.266
Veh	.607	.309	.583	.311	.663	**.353**	.663	.349
Wis	.924	.880	.924	.880	.924	.880	.924	.880
Wnq	.644	.344	.637	.352	.658	**.399**	.658	**.399**
Pim	.654	**.480**	.646	.477	.666	.460	.666	.460
Nty	.864	.720	.864	.720	.864	.720	.864	.720
Hab	.430	.055	.412	.053	.440	**.084**	.440	**.084**
Der	.989	.953	.989	.953	.989	.953	.989	.953
App	.636	.392	.636	.392	.636	.392	.636	.392
Mean	.721	.573	.713	.577	.729	**.583**	.729	**.583**

Tabla 5.2: Resultados en los conjuntos de entrenamiento (*Tr.*) y prueba (*Tst.*) obtenidos por las propuestas bipolares difusas con aprendizaje genético aplicadas al algoritmo CART con operador de agregación máx.

De la aplicación del método bipolar sobre el clasificador CART, Tabla 5.2, se pueden extraer las siguientes observaciones:

- Hay una mejora en kappa de 0.004 cuando se compara la agregación con la referencia seleccionada al azar y 0.01 en el caso de los enfoques bipolares tanto aditivos como logísticos.
- Los clasificadores bipolares aditivos y logísticos superan la clasificación de los enfoques restantes en 7 conjuntos de datos
- La referencia gana en 4 de los 25 conjuntos de datos y la agregación lo hace en 2 de ellos.
- Se producen empates en los resultados de 10 de los 25 conjuntos de datos.

En resumen, se consiguen mejoras o empates en 21 de los 25 conjuntos de datos considerados cuando se comparan los métodos bipolares frente a la referencia y en 19 de 25 al hacerlo también contra la agregación. Está claro que existe un comportamiento muy similar de ambos enfoques bipolares considerados cuando se aplican sobre el algoritmo CART.
En lo relativo a la aplicación del paradigma bipolar robusto sobre el clasificador RF, la Tabla 5.3 contiene los resultados alcanzados. En base a éstos, se destacan las siguientes particularidades:

- Hay una mejora en kappa medio de 0.006 cuando se compara el modelo bipolar aditivo con la referencia seleccionada al azar y 0.01 si se compara con la agregación difusa.
- El clasificador bipolar aditivo supera la clasificación de los enfoques restantes en 6 conjuntos de datos y el logístico lo hace en 9 de ellos.
- La referencia gana en 8 de 25 conjuntos de datos y la agregación lo hace en 2 de los conjuntos de datos considerados.
- Sa dan empates en los resultados de 3 de 25 conjuntos de datos.

Por lo tanto, podemos ver que se producen mejoras o empates en 17 de 25 conjuntos de datos al comparar solo con la referencia y en 15 de 25 de ellos al hacerlo también contra la agregación. Es importante tener en cuenta el comportamiento variable del método logístico bipolar en este caso. A pesar de ser el método ganador en varios conjuntos de datos, podemos ver que su media no es la mejor debido al menor valor kappa obtenido en varios de los conjuntos de datos restantes. Esto podría explicarse debido al carácter extremo de este tipo de función de agregación, como se explica en la Sección 4.1.
Teniendo en cuenta el clasificador ANN, el método bipolar alcanza los resultados mostrados en la Tabla 5.4 que podría interpretarse de la siguiente manera:

	RF Max							
	Ref		Aggr		bipAdd		bipLog	
	Tr.	*Tst*	*Tr.*	*Tst*	*Tr.*	*Tst*	*Tr.*	*Tst*
Car	1	.795	1	.801	1	.796	1	**.815**
Con	.707	.254	.664	.242	.804	.287	.807	**.292**
Eco	1	.549	1	.553	1	**.571**	1	**.571**
Fla	.583	.560	.597	.566	.643	**.622**	.653	.590
Gla	1	**.654**	1	**.654**	1	.645	1	.578
Lin	.959	**.744**	.975	.739	.991	.737	.991	.733
Shu	1	.996	1	.996	1	**.998**	1	.992
Thy	1	.886	1	.898	1	**.911**	1	.871
Yea	.999	.498	1	.507	1	.509	1	**.537**
Aus	.985	**.720**	.991	.718	.999	.717	.999	.713
Bal	.670	.575	.669	**.579**	.673	.577	.673	.577
Bup	1	**.469**	1	.461	1	.426	1	.442
Ger	1	.332	1	.338	1	.369	1	**.373**
Hay	1	1	1	1	1	1	1	.994
Pen	1	.951	1	.956	1	.957	1	**.972**
Pag	1	**.860**	1	.858	1	.853	1	.839
Sah	.971	.244	.971	.246	1	**.316**	1	.280
Veh	1	**.380**	1	.359	1	.365	1	.296
Wis	.994	**.941**	.995	.939	.997	.925	.997	.924
Wnq	1	.477	1	.476	1	.466	1	**.491**
Pim	1	**.506**	1	.501	1	.488	1	.493
Nty	1	.988	1	.988	1	.988	1	.974
Hab	.780	.162	.819	.074	.878	.150	.886	**.174**
Der	1	.985	1	.985	1	**1**	1	**1**
App	1	.551	1	.551	1	.551	1	.512
Mean	.946	.643	.947	.639	.959	**.649**	.960	.641

Tabla 5.3: Resultados en los conjuntos de entrenamiento (*Tr.*) y prueba (*Tst.*) obtenidos por las propuestas bipolares difusas con aprendizaje genético aplicadas al algoritmo RF con operador de agregación máx

- Hay una mejora en promedio de kappa de 0.002 al comparar la agregación con la referencia seleccionada al azar, siendo de 0.034 y 0.038 en el caso de enfoques bipolares tanto aditivos como logísticos, respectivamente.
- Los clasificadores bipolares aditivos y logísticos superan la clasificación de los enfoques restantes en 7 y 9 conjuntos de datos
- La referencia gana en 4 de los 25 conjuntos de datos y la agregación lo hace en 6 de ellos.
- Solo se da un empate en estos resultados.

Como resumen, se dan mejoras o empates en 21 de 25 conjuntos de datos

	ANN Max							
	Ref		Aggr		bipAdd		bipLog	
	Tr.	*Tst*	*Tr.*	*Tst*	*Tr.*	*Tst*	*Tr.*	*Tst*
Car	.998	.971	.999	.967	1	.966	1	**.972**
Con	.209	.226	.211	.218	.357	.327	.361	**.340**
Eco	.725	.488	.708	**.569**	.764	.514	.766	.466
Fla	.587	.614	.587	.610	.623	**.631**	.632	.606
Gla	.554	.454	.560	.454	.628	**.457**	.628	.452
Lin	.884	.778	.884	**.788**	.905	.774	.905	.779
Shu	.993	.974	.997	**.987**	.999	.983	.999	.982
Thy	.694	.632	.716	.594	.791	.626	.832	**.724**
Yea	.498	.477	.499	.477	.544	.547	.547	**.549**
Aus	.371	.310	.461	.455	.490	**.466**	.489	.459
Bal	.653	**.665**	.655	**.665**	.666	.641	.666	.641
Bup	.555	**.370**	.569	.310	.619	.358	.617	.357
Ger	.475	.360	.494	.414	.557	.422	.554	**.426**
Hay	.870	.726	1	.694	1	**.731**	1	.727
Pen	.988	.936	.991	**.938**	.995	.934	.996	.931
Pag	.927	.893	.928	.901	.941	**.905**	.945	.877
Sah	.381	**.338**	.476	.310	.532	.284	.532	.280
Veh	.546	.428	.634	.486	.691	.519	.691	**.525**
Wis	.946	.919	.947	**.928**	.957	.924	.958	.922
Wnq	.260	.281	.309	.269	.453	**.407**	.471	.402
Pim	.492	**.416**	.499	.355	.541	.405	.540	.399
Nty	.981	.964	.994	.949	1	**.973**	1	**.973**
Hab	.350	.219	.362	.188	.409	.192	.417	**.277**
Der	1	.985	1	.985	1	.985	1	.985
App	.165	.149	.037	.103	.528	.450	.674	**.481**
Mean	.644	.583	.661	.585	.720	.617	.729	**.621**

Tabla 5.4: Resultados en los conjuntos de entrenamiento (*Tr.*) y prueba (*Tst.*) obtenidos por las propuestas bipolares difusas con aprendizaje genético aplicadas al algoritmo ANN con operador de agregación máx

al comparar solo con la referencia y en 15 de 25 al hacerlo también contra la agregación.

5.2.2.2. Operador de agregación *Mínimo*

Bajo la perspectiva de agregación mediante el operador mínimo, las Tablas 5.5,5.6 y 5.7 muestran un buen comportamiento a nivel general del método de toma de decisiones bipolar, al menos bajo uno de los paradigmas de agregación bipolar. debido a que se mejora el rendimiento de los clasificadores base y difuso agregado.

De la aplicación del paradigma bipolar al algoritmo CART, Tabla 5.5, se extraen las siguientes observaciones:

	CART Min							
	Ref		Aggr		bipAdd		bipLog	
	Tr.	*Tst*	*Tr.*	*Tst*	*Tr.*	*Tst*	*Tr.*	*Tst*
Car	.796	.727	.795	.733	.797	**0.734**	.797	**.734**
Con	.391	.247	.336	.232	.371	.298	.401	**.325**
Eco	.800	.642	.791	.595	.791	.595	.793	.595
Fla	.620	.560	.604	.559	.627	**.561**	.627	**.561**
Gla	.722	.488	.720	.473	.722	**.513**	.723	.489
Lin	.667	.620	.667	.620	.667	.620	.667	.620
Shu	.994	.995	.994	.995	.994	.995	.994	.995
Thy	.876	**.819**	.876	**.819**	.879	.806	.879	.806
Yea	.605	.474	.574	.480	.585	.488	.594	**.498**
Aus	.785	.673	.782	.684	.795	**.691**	.795	**.691**
Bal	.655	.518	.655	.518	.655	.518	.655	.518
Bup	.623	.346	.618	.344	.621	**.349**	.621	**.349**
Ger	.503	.301	.518	.335	.538	**.342**	.538	**.342**
Hay	.841	.738	.841	.738	.841	.738	.841	.738
Pen	.970	.932	.970	.932	.970	.932	.970	.932
Pag	.872	.887	.872	.887	.872	.887	.872	.887
Sah	.545	.220	.511	**.340**	.578	.258	.582	.266
Veh	.607	.309	.583	.311	.629	.344	.651	**.352**
Wis	.924	.880	.924	.880	.924	.880	.924	.880
Wnq	.644	.344	.637	.352	.644	**.408**	.645	.407
Pim	.654	**.480**	.646	.477	.657	.456	.662	.477
Nty	.864	.720	.864	.720	.864	.720	.864	.720
Hab	.430	.055	.412	.053	.422	.082	.438	**.136**
Der	.989	.953	.989	.953	.989	.953	.989	.953
App	.636	.392	.636	.392	.636	.392	.636	.392
Mean	.721	.573	.713	.577	.723	.582	.726	**.586**

Tabla 5.5: Resultados en los conjuntos de entrenamiento (*Tr.*) y prueba (*Tst.*) obtenidos por las propuestas bipolares difusas con aprendizaje genético aplicadas al algoritmo CART con operador de agregación mín.

- Existe una mejora en kappa de 0.004 cuando se compara la agregación con un clasificador probabilístico seleccionado al azar y 0.013 en caso del enfoque logístico bipolar.
- El modelo bipolar aditivo mejora los resultados de los clasificadores restantes en 7 conjuntos de datos y el logístico lo hace en 9 de ellos.
- La referencia y la agregación difusa ganan en 2 de 25 conjuntos de datos cada uno.
- Se producen empates en los resultados de 11 de los 25 conjuntos de datos.

En resumen, hemos logrado mejoras o empates en 23 de 25 conjuntos de

datos al comparar solo con la referencia y en 22 de 25 de ellos al hacerlo también contra la agregación. Está claro que el enfoque logístico bipolar presenta un mejor patrón de comportamiento que el modelo aditivo cuando se aplican al algoritmo CART.

	RF Min							
	Ref		Aggr		bipAdd		bipLog	
	Tr.	*Tst*	*Tr.*	*Tst*	*Tr.*	*Tst*	*Tr.*	*Tst*
Car	1	.795	1	.801	1	.795	1	**.825**
Con	.707	.254	.664	.242	.787	**.304**	.801	.297
Eco	1	.549	1	.553	1	**.586**	1	.571
Fla	.583	.560	.597	.566	.630	**.624**	.648	.615
Gla	1	**.654**	1	**.654**	1	.640	1	.614
Lin	.959	**.744**	.975	.739	.991	.711	.991	.703
Shu	1	.996	1	.996	1	**.998**	1	.996
Thy	1	.886	1	**.898**	1	.892	1	.843
Yea	.999	.498	1	.507	1	.509	1	**.516**
Aus	.985	**.720**	.991	.718	.999	.719	.999	.718
Bal	.670	.575	.669	**.579**	.673	.577	.673	.577
Bup	1	**.469**	1	.461	1	.412	1	.423
Ger	1	.332	1	.338	1	.344	1	**.401**
Hay	1	1	1	1	1	1	1	1
Pen	1	.951	1	.956	1	.958	1	**.962**
Pag	1	**.860**	1	.858	1	.853	1	.835
Sah	.971	.244	.971	.246	1	**.305**	1	.279
Veh	1	.380	1	.359	1	.340	1	**.389**
Wis	.994	**.941**	.995	.939	.997	.935	.997	.931
Wnq	1	**.477**	1	.476	1	.473	1	.475
Pim	1	.506	1	.501	1	**.509**	1	.470
Nty	1	**.988**	1	**.988**	1	.974	1	.822
Hab	.780	.162	.819	.074	.867	.132	.884	**.171**
Der	1	.985	1	.985	1	**.988**	1	.985
App	1	.551	1	.551	1	.551	1	.551
Mean	.946	.643	.947	.639	.958	**.645**	.960	.639

Tabla 5.6: Resultados en los conjuntos de entrenamiento (*Tr.*) y prueba (*Tst.*) obtenidos por las propuestas bipolares difusas con aprendizaje genético aplicadas al algoritmo RF con operador de agregación mín.

En la aplicación de este esquema bipolar al clasificador RF, se obtienen los resultados contenidos en la Tabla 5.6, que proporciona las siguientes observaciones:

- Hay una mejora por medio de kappa de 0.002 cuando se compara el modelo bipolar aditivo con la referencia seleccionada al azar y 0.06 en caso de compararlo con la agregación difusa.
- El clasificador bipolar aditivo supera la clasificación de los enfoques

restantes en 7 conjuntos de datos y el bipolar lo hace en 6 de ellos.

- La referencia gana en 8 de 25 conjuntos de datos y la agregación lo hace en 4 de los conjuntos de datos considerados.
- Hay un único empate en estos resultados.

Por lo tanto, hemos alcanzado mejoras o vínculos en 17 de 25 conjuntos de datos al comparar solo con la referencia y en 15 de 25 de ellos al hacerlo también contra la agregación. Una vez más, debemos tener en cuenta el comportamiento extraño de la metodología, ya sea en la agregación difusa o en los enfoques logísticos bipolares.

	ANN Min							
	Ref		Aggr		bipAdd		bipLog	
	Tr.	*Tst*	*Tr.*	*Tst*	*Tr.*	*Tst*	*Tr.*	*Tst*
Car	.998	**.971**	.999	.967	1	.968	1	.967
Con	.209	.226	.211	.218	.352	.324	.361	**.333**
Eco	.725	.488	.708	**.569**	.760	.512	.762	.510
Fla	.587	.614	.587	.610	.621	**.630**	.632	.605
Gla	.554	.454	.560	.454	.630	.484	.627	**.486**
Lin	.884	.778	.884	**.788**	.905	.761	.905	.761
Shu	.993	.974	.997	**.987**	.999	**.987**	.999	.982
Thy	.694	.632	.716	.594	.792	.663	.890	**.872**
Yea	.498	.477	.499	.477	.543	.545	.547	**.555**
Aus	.371	.310	.461	.455	.490	**.467**	.490	**.467**
Bal	.653	**.665**	.655	**.665**	.666	.641	.666	.641
Bup	.555	**.370**	.569	.310	.608	.355	.611	.351
Ger	.475	.360	.494	.414	.548	**.429**	.548	.415
Hay	.870	**.726**	1	.694	1	.694	1	.694
Pen	.988	.936	.991	.938	.991	.938	.994	**.949**
Pag	.927	.893	.928	**.901**	.940	.876	.944	.876
Sah	.381	**.338**	.476	.310	.527	.286	.531	.284
Veh	.546	.428	.634	.486	.694	.538	.705	**.539**
Wis	.946	.919	.947	.928	.958	**.929**	.958	**.929**
Wnq	.260	.281	.309	.269	.455	**.395**	.467	.371
Pim	.492	.416	.499	.355	.535	**.439**	.537	.409
Nty	.981	.964	.994	.949	1	**.973**	1	**.973**
Hab	.350	.219	.362	.188	.425	.210	.422	**.224**
Der	1	.985	1	.985	1	.985	1	.985
App	.165	.149	.037	.103	.265	.156	.643	**.462**
Mean	.644	.583	.661	.585	.708	.607	.729	**.626**

Tabla 5.7: Resultados en los conjuntos de entrenamiento (*Tr.*) y prueba (*Tst.*) obtenidos por las propuestas bipolares difusas con aprendizaje genético aplicadas al algoritmo ANN con operador de agregación mín.

Teniendo en cuenta el clasificador ANN, el método bipolar alcanza los

resultados que se muestran en la Tabla 5.7, que pueden ser interpretados como sigue:

- Se produce una mejora en kappa de 0.002 al comparar la agregación con la referencia seleccionada al azar, siendo de 0.043 y 0.024 en el caso de enfoques bipolares tanto aditivos como logísticos, respectivamente.
- Los clasificadores bipolares aditivos y logísticos superan la clasificación de los enfoques restantes en 8 y 11 conjuntos de datos.
- La referencia gana en 4 de 25 conjuntos de datos y la agregación lo hace en 5 de ellos.
- Solo hay un empate en este resultado.

En general, hemos alcanzado mejoras o empates en 21 de 25 conjuntos de datos al comparar solo con la referencia y en 16 de 25 de ellos al hacerlo también contra la agregación.

5.2.2.3. Operador de agregación *Media Aritmética*

Cuando el método bipolar se aplica al clasificador CART por medio de la consideración de una agregación difusa bajo el operador *media aritmética*, Tabla 5.8, los resultados podrían interpretarse de la siguiente manera:

- Hay una mejora en kappa de 0.008 al comparar los modelos bipolares aditivo y logístico con la referencia seleccionada al azar y 0.01 en caso de compararlos con el clasificador agregado difuso.
- Los clasificadores bipolares tanto aditivos como logísticos superan la clasificación de los enfoques restantes en 6 y 7 conjuntos de datos.
- La referencia y la agregación obtienen resultados superiores en 3 de 25 conjuntos de datos cada uno.
- Se producen empates en 12 de 25 conjuntos de datos.

En consecuencia, podemos ver que hemos alcanzado mejoras o empates en 22 de 25 conjuntos de datos al comparar solo con la referencia y en 20 de 25 de ellos al hacerlo también contra la agregación. Es claro que existe un comportamiento muy similar de ambos enfoques bipolares considerados cuando se aplican al algoritmo CART.

En el caso del método bipolar aplicado al clasificador de RF, en la Tabla 5.9 se muestran los resultados que dan pie a una breve descripción de su comportamiento en la forma siguiente:

	CART							
	Mean							
	Ref		Aggr		bipAdd		bipLog	
	Tr.	*Tst*	*Tr.*	*Tst*	*Tr.*	*Tst*	*Tr.*	*Tst*
Car	.796	.727	.795	.733	.797	.733	.797	.733
Con	.391	.247	.379	.245	.409	.284	.410	**.301**
Eco	.800	.642	.793	.595	.797	.642	.797	.642
Fla	.620	.560	.620	.554	.627	**.561**	.627	**.561**
Gla	.722	.488	.722	.454	.723	**.489**	.723	**.489**
Lin	.667	.620	.667	.620	.667	.620	.667	.620
Shu	.994	.995	.994	.995	.994	.995	.994	.995
Thy	.876	**.819**	.876	**.819**	.879	.806	.879	.806
Yea	.605	.474	.584	**.484**	.614	.478	.614	.478
Aus	.785	.673	.782	.666	.790	**.680**	.790	**.680**
Bal	.655	.518	.655	.518	.655	.518	.655	.518
Bup	.623	.346	.623	.346	.623	.346	.623	.346
Ger	.503	.301	.485	**.352**	.511	.348	.511	.348
Hay	.841	.738	.841	.738	.841	.738	.841	.738
Pen	.970	.932	.970	.932	.970	.932	.972	**.939**
Pag	.872	.887	.872	.887	.872	.887	.872	.887
Sah	.545	**.220**	.572	.217	.584	.215	.584	.215
Veh	.607	.309	.627	.302	.656	**.352**	.662	.346
Wis	.924	.880	.924	.880	.924	.880	.924	.880
Wnq	.644	.344	.646	.337	.659	**.393**	.659	**.393**
Pim	.654	**.480**	.646	.477	.665	.470	.667	.467
Nty	.864	.720	.864	.720	.864	.720	.864	.720
Hab	.430	.055	.430	.055	.440	**.084**	.440	**.084**
Der	.989	.953	.989	.953	.989	.953	.989	.953
App	.636	.392	.636	.392	.636	.392	.636	.392
Mean	.721	.573	.720	.571	.728	**.581**	.728	**.581**

Tabla 5.8: Resultados en los conjuntos de entrenamiento (*Tr.*) y prueba (*Tst.*) obtenidos por las propuestas bipolares difusas con aprendizaje genético aplicadas al algoritmo CART con operador de agregación *media aritmética*.

- Exite una mejora en términos de kappa de 0.004 cuando se compara el modelo bipolar logístico con la referencia seleccionada al azar y 0.007 en caso de compararlo con la agregación.
- El clasificador bipolar aditivo supera la clasificación de los enfoques restantes en 6 conjuntos de datos y el logístico lo hace en 8 de ellos.
- La referencia obtienen mejores resultados en 6 de 25 conjuntos de datos y la agregación lo hace en 4 de ellos los conjuntos de datos considerados.
- Se producen empates en los resultados de 3 de 25 conjuntos de datos.

En resumen, hemos alcanzado mejoras o empates en 19 de 25 conjuntos de datos al comparar solo con la referencia y en 15 de 25 de ellos al hacerlo

	RF Mean							
	Ref		Aggr		bipAdd		bipLog	
	Tr.	*Tst*	*Tr.*	*Tst*	*Tr.*	*Tst*	*Tr.*	*Tst*
Car	1	.795	1	.801	1	.799	1	**.810**
Con	.707	.254	.663	.244	.799	**.299**	.805	.295
Eco	1	.549	1	.553	1	.570	1	**.577**
Fla	.583	.560	.584	.571	.631	**.617**	.650	.611
Gla	1	.654	1	**.656**	1	.611	1	.683
Lin	.959	**.744**	.975	.739	.991	.703	.991	.717
Shu	1	.996	1	**.998**	1	**.998**	1	.996
Thy	1	.886	1	**.898**	1	.863	1	.893
Yea	.999	.498	1	.507	1	.519	1	**.528**
Aus	.985	.720	.989	.719	.995	**.731**	.995	.724
Bal	.670	.575	.671	.569	.673	**.580**	.673	**.580**
Bup	1	**.469**	1	.450	1	.446	1	.401
Ger	1	.332	1	**.335**	1	.275	1	.318
Hay	1	1	1	1	1	1	1	1
Pen	1	.951	1	.956	1	**.959**	1	**.959**
Pag	1	**.860**	1	.853	1	.862	1	.835
Sah	.971	.244	.974	.253	1	.291	1	**.322**
Veh	1	**.380**	1	.365	1	.314	1	.362
Wis	.994	**.941**	.995	.939	.997	.934	.997	.922
Wnq	1	.477	1	.472	1	**.489**	1	.476
Pim	1	**.506**	1	.500	1	.424	1	.460
Nty	1	.988	1	.988	1	.958	1	.988
Hab	.780	.162	.800	.103	.870	.168	.880	**.177**
Der	1	.985	1	.985	1	.985	1	**1**
App	1	.551	1	.551	1	.551	1	.551
Mean	.946	.643	.946	.640	.958	.638	.960	**.647**

Tabla 5.9: Resultados en los conjuntos de entrenamiento (*Tr.*) y prueba (*Tst.*) obtenidos por las propuestas bipolares difusas con aprendizaje genético aplicadas al algoritmo ANN con operador de agregación *media aritmética*.

también contra la agregación. Parece que el clasificador bipolar aditivo presenta los peores resultados por el bajo valor kappa obtenido en conjuntos de datos como *German* y *Vehicle*.

Tomando en consideración el clasificador ANN, el método bipolar bajo agregación mediante la *media aritmética* obtiene los resultados que se muestran en la Tabla 5.10, de donde se extraen las siguientes características sobre su comportamiento:

- Hay una mejora por medio de kappa de 0.041 cuando se compara el modelo bipolar logístico con la referencia seleccionada al azar, siendo 0.047 en caso de hacerlo contra el clasificador agregado difuso.
- El clasificador logístico bipolar supera la clasificación de los enfoques

	ANN Mean							
	Ref		Aggr		bipAdd		bipLog	
	Tr.	*Tst*	*Tr.*	*Tst*	*Tr.*	*Tst*	*Tr.*	*Tst*
Car	.998	**.971**	.999	.967	1	.968	1	**.971**
Con	.209	.226	.210	.215	.355	.316	.364	**.336**
Eco	.725	.488	.718	**.557**	.760	.514	.763	.514
Fla	.587	.614	.589	.619	.622	**.627**	.632	.606
Gla	.554	.454	.561	.456	.638	**.519**	.640	.506
Lin	.884	.778	.884	.788	.905	.761	.905	**.779**
Shu	.993	.974	.999	**.986**	.999	.981	.999	.984
Thy	.694	.632	.667	.540	.787	.675	.858	**.785**
Yea	.498	.477	.498	.477	.543	.545	.547	**.548**
Aus	.371	.310	.441	.396	.480	.419	.480	**.420**
Bal	.653	**.665**	.654	**.665**	.665	.641	.666	.641
Bup	.555	.370	.580	.326	.614	.382	.612	**.384**
Ger	.475	.360	.502	.424	.551	**.438**	.550	.418
Hay	.870	**.726**	.953	.694	.961	.720	.961	.720
Pen	.988	.936	.991	**.943**	.995	.937	.996	**.943**
Pag	.927	.893	.922	**.901**	.939	.868	.946	.880
Sah	.381	**.338**	.462	.312	.509	.288	.511	.266
Veh	.546	.428	.629	.489	.695	.555	.693	**.558**
Wis	.946	.919	.948	**.930**	.958	.918	.958	.918
Wnq	.260	.281	.269	.254	.440	.384	.455	**.410**
Pim	.492	.416	.508	.380	.544	.410	.545	**.425**
Nty	.981	.964	.994	.949	1	**.973**	1	**.973**
Hab	.350	.219	.365	.184	.422	**.235**	.422	.216
Der	1	.985	1	.985	1	.985	1	**1**
App	.165	.149	.033	.000	.253	.156	.625	**.389**
Mean	.644	.583	.655	.577	.705	.609	.725	**.624**

Tabla 5.10: Resultados en los conjuntos de entrenamiento (*Tr.*) y prueba (*Tst.*) obtenidos por las propuestas bipolares difusas con aprendizaje genético aplicadas al algoritmo ANN con operador de agregación *media aritmética*.

de resto en 14 conjuntos de datos.

- La referencia gana en 4 de los 25 conjuntos de datos y la agregación lo hace en 6 de ellos.

- No se produce empate alguno en estos resultados.

Consiguientemente, se producen mejoras o empates en 21 de 25 conjuntos de datos al comparar solo con la referencia y en 16 de 25 de ellos al hacerlo también contra la agregación.

5.2.3. Análisis estadístico

Con el fin de detectar diferencias significativas entre los resultados de los diferentes enfoques, se aplica en este punto el esquema de pruebas estadísticas adecuadas para las comparaciones en rendimiento de algoritmos en el contexto de la MDD [18, 33]. En concreto, se lleva a cabo el test de rangos alineados de Friedman para evaluar los resultados a nivel general, esto es, teniendo en cuenta un marco de comparaciones múltiples. Para las comparaciones por pares de métodos se utiliza el test de Wilcoxon. Procediendo de esta forma se alcanzan, para cada operador de agregación, los resultados detallados en las siguientes secciones. En todos los casos se representa con * el nivel de significación menor al 5 %.

5.2.3.1. Operador de agregación *Máximo*

En el caso de considerar el operador máximo, el test de Friedman obtiene un p-valor bajo para los tres algoritmos, lo que implica que existen diferencias significativas entre los resultados proporcionados por cada método.

Por este motivo, podemos aplicar una prueba post hoc para comparar nuestra metodología con los enfoques restantes. Específicamente, se aplica una prueba Holm utilizando el mejor método (aquel con menor rango) como método de control y calculando el p-valor ajustado (APV) para los métodos restantes.

Algoritmo	Rango CART	Rango RF	Rango ANN
Ref"	62.2	52.54	62.14
.^{A}ggr"	58.76	55.3	59.98
"BipAdd"	41.56	42.36	39.42
"BipLog"	39.48	51.8	40.46
p-val	.000163	.000135	.000147
APV Ref"	**.0168***	**.4295**	**.0168***
APV .^{A}ggr"	**.0375***	**.3444**	**.0244***

Tabla 5.11: Rangos promedio (Aligned Friedman), p-kappaes asociados y APV del test de Holm para cada algoritmo. Agregación máx.

El test post-hoc de Holm, como se muestra en la Tabla 5.11, refleja que el método bipolar supera a los clasificadores de referencia y agregados con un alto nivel de confianza en el caso del algoritmo CART, en el que el enfoque logístico parece ser el mejor, y el clasificador ANN donde el aditivo alcanza el menor rango (es decir, resulta el mejor).

Con respecto al clasificador de RF base, debido al extraño comportamiento de nuestro método logístico bipolar cuando se aplica a este clasificador, la prueba post-hoc de Holm indica que no hay suficiente evidencia estadística

para afirmar que nuestro método logra mejores resultados que los clasificadores de referencia o agregado. Sin embargo, el modelo bipolar aditivo alcanza el rango más bajo con una diferencia apreciable.

Comparación	R^+	R^-	p-val
CARTbipAdd vs. CARTRef	213.5	86.5	.0674
CARTbipLog vs. CARTRef	236.5	88.5	**.0450**
CARTbipAdd vs. CARTAggr	224.5	100.5	.0926
CARTbipLog vs. CARTAggr	214.0	86.0	.0653
ANNbipAdd vs. ANNRef	215.0	85.0	.0612
ANNbipLog vs. ANNRef	206.0	94.0	.1064
ANNbipAdd vs. ANNAggr	228.0	72.0	**.0249**
ANNbipLog vs. ANNAggr	216.0	84.0	.0574

Tabla 5.12: Test de Wilcoxon para comparar los métodos bipolares (R^+) frente al clasificador base (R^-). Agregación máx.

El análisis estadístico de las comparaciones por pares de métodos, que se realiza mediante una prueba de Wilcoxon, Tabla 5.12, refleja la superioridad de la metodología propuesta cuando se aplica a los algoritmos CART y ANN con p-kappaes aceptables, especialmente cuando se compara el modelo bipolar logístico con la referencia para el algoritmo CART y el aditivo con el clasificador agregado en el caso de ANN. Nuevamente, la aplicación de la metodología en el algoritmo de RF no alcanza mejoras significativas.

En general, considerando el máximo como operador de agregación para obtener el clasificador difuso, hemos alcanzado buenos resultados para CART y ANN, no siendo el caso cuando se trata de RF.

5.2.3.2. Operador de agregación*Mínimo*

Para la función de agregación mínimo, el test de rangos alineados de Friedman obtiene un p-valor bajo para los tres algoritmos, véase la Tabla 5.13, lo que implica que hay diferencias significativas entre los resultados.

Algoritmo	Rango CART	Rango RF	Rango ANN
Ref"	63.34	49.72	59.98
.Aggr"	57.86	53.88	59.92
"BipAdd"	41.56	46.86	43.58
"BipLog"	39.24	51.54	38.52
p-val	.000156	.000133	.000161
Holm Ref"	**.0099***	**1**	**.0267***
Holm .Aggr"	**.0465***	**1**	**.0267***

Tabla 5.13: Rangos promedio de los algoritmos (Aligned Friedman), p-valores asociados y p-valor Ajustado de Holm para cada algoritmo. Agregación mín.

Por este motivo, podemos aplicar una prueba post hoc para comparar nuestra metodología con los enfoques restantes. Esta prueba, como se muestra en la Tabla 5.13, refleja que el método bipolar supera a los clasificadores de referencia y agregados con un alto nivel de confianza en el caso de los algoritmos CART y ANN, en los que el enfoque logístico parece ser el mejor.

Contrariamente, y debido al extraño comportamiento del enfoque logístico cuando se aplica al algoritmo de RF, el test post-hoc de Holm indica que no hay evidencia estadística suficiente para afirmar que nuestro método logra mejores resultados que los clasificadores de referencia o agregados. Sin embargo, el modelo bipolar aditivo alcanza nuevamente la clasificación más baja.

Comparación	R^+	R^-	p-val
CARTbipAdd vs. CARTRef	243.5	81.5	**.0283**
CARTbipLog vs. CARTRef	246.5	78.5	**.0229**
CARTbipAdd vs. CARTAggr	213.5	86.5	.0674
CARTbipLog vs. CARTAggr	233.5	66.5	**.0163**
ANNbipAdd vs. ANNRef	217.0	83.0	.0537
ANNbipLog vs. ANNRef	216.0	84.0	.0574
ANNbipAdd vs. ANNAggr	231.5	68.5	**.0191**
ANNbipLog vs. ANNAggr	241.5	83.5	**.0324**

Tabla 5.14: Test de Wilcoxon para comparar los métodos bipolares (R^+) frente al clasificador base (R^-). Agregación mín

El test de Wilcoxon, Tabla 5.14, refleja la superioridad de nuestra nueva metodología cuando se aplica a los algoritmos CART y ANN con p-valores muy bajos, especialmente cuando se comparan los dos modelos bipolares con la referencia y el logístico con el agregación difusa para el algoritmo CART y los dos modelos bipolares contra el clasificador agregado en el caso de ANN. Nuevamente, la aplicación de la metodología en el algoritmo de RF no alcanza mejoras significativas.

Con todo, considerando el mínimo como operador de agregación para obtener el clasificador difuso, hemos alcanzado resultados positivos para CART y ANN, no así en el caso de RF.

5.2.3.3. Operador de agregación *Media Aritmética*

Una vez más, el test de rangos alineados de Friedman obtiene un p-valor bajo para los tres algoritmos, por lo que hay diferencias significativas entre los resultados. En consecuencia, se aplica el test post hoc de Holm, que se muestra en la Tabla 5.15, en la que se pone de manifiesto la superioridad del método bipolar con respecto tanto a la referencia como a los clasificadores agregados con un nivel de confianza aceptable en el caso de CART y ANN, donde el clasificador logístico bipolar logra el rango más bajo.

Algoritmo	Rango CART	Rango RF	Rango ANN
Ref"	58.2	51.54	63.56
.^{A}ggr"	60.08	54.74	60.88
"BipAdd"	43	53.3	41.32
"BipLog"	40.72	42.42	36.24
p-val	.000156	.000130	.000142
Holm Ref"	**.0663**	**.399**	**.0026***
Holm .^{A}ggr"	**.0549**	**.399**	**.0053***

Tabla 5.15: Rangos promedio de los algoritmos (Aligned Friedman), p-valores asociados y p-valor Ajustado de Holm para cada algoritmo. Agregación *media aritmética*.

De manera opuesta y dado el extraño comportamiento de nuestro enfoque bipolar aditivo cuando se aplica al algoritmo de RF con la agregación *media aritmética*, el test post-hoc de Holm indica que no hay suficiente evidencia estadística para afirmar que nuestro método logra mejores resultados que el clasificador de referencia o agregado. Sin embargo, el modelo logístico bipolar alcanza la clasificación más baja con una diferencia apreciable.

Comparación	R^+	R^-	p-val
CARTbipAdd vs. CARTRef	231.0	94.0	.0633
CARTbipLog vs. CARTRef	221.5	78.5	**.0396**
CARTbipAdd vs. CARTAggr	201.5	98.5	.137
CARTbipLog vs. CARTAggr	223.5	101.5	.0979
ANNbipAdd vs. ANNRef	231.0	69.0	**.0198**
ANNbipLog vs. ANNRef	262.0	63.0	**.0071**
ANNbipAdd vs. ANNAggr	231.0	69.0	**.0198**
ANNbipLog vs. ANNAggr	249.0	76.0	**.0192**

Tabla 5.16: Test de Wilcoxon para comparar los métodos bipolares (R^+) frente al clasificador base (R^-). Agregación *media aritmética*.

Las comparaciones por pares realizadas por una prueba de Wilcoxon, Tabla 5.16, reflejan la mejora de nuestra nueva metodología cuando se aplica a los algoritmos CART y ANN con p-valores aceptables, especialmente cuando se compara el modelo logístico bipolar con el referencia para el algoritmo CART y para todos los pares en caso de ANN. Nuevamente, la aplicación de la metodología sobre el algoritmo de RF no alcanza mejoras significativas.

5.2.4. Principales conclusiones

En primer lugar, de la evaluación del comportamiento de los tres operadores de agregación (máximo, mínimo y medio) utilizados para obtener los clasificadores difusos, se puede concluir que se han logrado resultados muy positivos para CART y ANN y no tan buenos en el caso de RF. El análisis

completo en este sentido contiene las siguientes aseveraciones:

- Considerando los clasificadores básicos CART y ANN, el clasificador combinado difuso mejora ligeramente el comportamiento de la referencia en promedio de kappa cuando se utilizan los operadores de agregación máximo y mínimo. No es así en el caso de la agregación por la media, donde el clasificador difuso muestra una ligera disminución de la media de kappa. Con respecto al algoritmo de RF, no se alcanza ninguna mejora en esta comparación para los tres operadores de agregación.

- Tanto los clasificadores bipolares aditivo como los logístico superan la clasificación de los restantes enfoques para los tres operadores de agregación considerados cuando se aplican a los algoritmos CART y ANN. En el caso de RF, el modelo bipolar aditivo es el que tiene la mejor media para las agregaciones máxima y mínima, y el logístico obtiene el mejor resultado para la agregación por el operador de la media

- El test de rango alineado de Friedman refleja que hay diferencias estadísticas entre los cuatro enfoques en todas las comparaciones múltiples para las tres agregaciones aplicadas a los tres algoritmos. Sin embargo, la prueba post hoc de Holm aplicada al algoritmo de RF señala que no hay pruebas suficientes para afirmar que el método de control (bipolar aditivo en caso de máximo y mínimo y bipolar logístico para la agregación de la media) mejora los resultados alcanzados por cualquiera la referencia o su agregación difusa.

- Con respecto a las comparaciones por pares, el test de rango de Wilcoxon señala que los modelos logísticos aplicados al algoritmo de ANN mejoran claramente el comportamiento del clasificador difuso cuando se realiza mediante operadores mínimos y medios, y solo lo hacen en el modelo aditivo para el máximo . En el caso del algoritmo CART podemos ver mejoras estadísticamente significativas para los siguientes pares; bipolar logístico contra la referencia para operadores máximos y medios, y en todos menos agregación bipolar vs. difusa aditiva cuando se enfoca en el mínimo.

- Desafortunadamente, no hemos alcanzado mejoras estadísticamente significativas en los resultados para el clasificador de RF. Sin embargo, podemos ver algunas mejoras interesantes en varios conjuntos de datos, especialmente en el caso de la agregación de medias.

La aplicación de una estructura de disimilitud a la evidencia *soft* proporcionada por los clasificadores difusos conduce a un marco de representación

bipolar del conocimiento en el contexto difuso. Para estudiar la viabilidad del enfoque propuesto, y en particular para señalar que es susceptible de aplicación sobre cualquier clasificador difuso a pesar de cómo se obtiene, lo hemos aplicado a un conjunto de clasificadores difusos robustos obtenidos mediante la agregación de las estimaciones de distribución de probabilidad replicadas proporcionadas por diferentes clasificadores probabilísticos de base, específicamente árboles de clasificación CART, Random Forest y Redes Neuronales.

En resumen, este análisis proporciona las siguientes aseveraciones finales:

- El marco bipolar permitió mejorar significativamente los resultados de los tres algoritmos de aprendizaje de máquina base considerados en este trabajo.
- Tanto el método aditivo como el de ajuste logístico superan significativamente los resultados del clasificador base en el caso de CART y ANN, pero solo el método aditivo mejora el comportamiento de RF en términos estadísticos.
- Comparando tanto el aditivo como el clasificador logístico propuesto, encontramos que no hay un ganador claro. De hecho, esta pregunta parece depender de alguna manera del algoritmo base considerado, así como del conjunto de datos de la aplicación.

En base a estos resultados, el enfoque propuesto parece proporcionar una solución factible para enfrentar los problemas de clasificación binaria y mejorar en algunos casos la regla de decisión que administra cómo se explota la información flexible intermedia recopilada por diversos clasificadores.

Capítulo 6

Nuevo método local en el contexto de los Sistemas de Clasificación Basados en Reglas Difusas (SCBRDs)

El lenguaje ha creado al hombre más que el hombre al lenguaje.

Jacques Monod

Resumen: A lo largo de este capítulo, se propone y evalúa una novedosa metodología para la Representación Bipolar del Conocimiento (RBC) en un nivel local o nivel de reglas, en este caso, generadas por cualquier Sistema de Clasificación Basado en Reglas Difusas (SCBRD). Se trata de una particularización del marco general de RBC presentado en el Capítulo 3, esta vez, en el contexto de los SCBRDs. El objetivo fundamental de esta metodología es introducir un marco de representación del conocimiento que que permita una mejor adaptación de los SCBRDs a las particularidades de ciertas regiones del espacio de características, generando pares de evidencia bipolar a nivel de regla. En la Sección 6.1 se presenta el desarrollo teórico de esta propuesta así como la definición de los conceptos de excepciones mayores y menores a una regla de clasificación (Sección 6.1.3) y una propuesta para la extensión bipolar del Método de Razonamiento Difuso presentado en 2.6.3 (Sección 6.1.4). Los resultados de la aplicación de esta metodología se encuentran recogidos en la Sección 6.2, con la configuración experimental (6.2), un caso de estudio teórico (6.2.2) y un experimento con datos reales (6.2.3).

6.1. Representación Bipolar del Conocimiento a nivel local o de reglas en SCBRDs

Esta sección está enteramente dedicada a la presentación de una de las aportaciones más relevantes de este trabajo. Una vez han sido exploradas diversos métodos bipolares concebidos para su aplicación en un *Nivel Global*, ya sea en entornos probabilísticos o difusos, en el que la RBC toma en consideración la evidencia *soft* dada por el clasificador en la etapa previa a la toma de decisiones final se pueden valorar sus bondades y defectos. En primer lugar, esta propuesta tiene la gran ventaja de ser susceptible de aplicación sobre **cualquier** algoritmo de clasificación de naturaleza *soft*, lo que hace de la flexibilidad su mayor fortaleza. No obstante, el paradigma a *Nivel Global* no tiene sensibilidad local y, por ello, encuentra diversos inconvenientes en su desarrollo como lo es, por ejemplo, el problema de aplicación sobre clasificadores de muy alto rendimiento discutido en la Sección 3.3.2.

Debido a esto, y con el objetivo de soslayar algunos de los inconvenientes expuestos, se apuesta ahora por una aplicación del paradigma bipolar en lo que hemos llamado *Nivel Local* o *Nivel de Reglas*.

El método local propuesto introduce una representación bipolar en el nivel de regla de los SCBRDs, con el objetivo fundamental de evaluar el alcance de una posible sinergia entre dos de los métodos de representación de conocimiento inspirados en el comportamiento humano. Por un lado, los SCBRDs están basados en la forma de comunicación humana por medio de términos lingüísticos que manejan adecuadamente conceptos de naturaleza imprecisa. De otra parte, la RBC se inspira en la forma en la que el ser humano razona y toma decisiones, frecuentemente en base a afectos de carácter dicotómico (positivos y negativos). Por ello, es razonable pensar que la unión de tales paradigmas en un único modelo de clasificación proporcionará una representación más cercana a nuestro entendimiento y, en el mejor de los casos, también una clasificación con un mayor rendimiento.

Se aplican en esta sección dos enfoques bipolares diferentes en el contexto de SCBRDs. Por un lado, el enfoque global "a posteriori" presentado en las Secciones 4.1 y 5.1, en contextos probabilístico y difuso, respectivamente y que es concebido para su aplicación en el último paso de clasificación del proceso de inferencia (señalado en verde en la Figura 6.1). Por el otro, un nuevo enfoque local, concebido para ser aplicado en la base de reglas después de entrenar el algoritmo (subrayado en azul en la Figura 6.1). Por lo tanto, queda claro que el nuevo enfoque local trabaja a un nivel más profundo de los SCBRDs, aprovechando la información flexible contenida en la base de reglas (BR) de la primera etapa del proceso de clasificación.

Esta sección está organizada de la siguiente manera. En primer lugar se presenta en la Sección 6.1.1 una particularización del marco general de bipolaridad global en contexto difuso propuesto en la Sección 5.1, esta vez,

considerando como evidencias los grados de consistencia o solidez dados por el SCBRD. El método local propuesto se describe en la Sección 6.1.2, definiendo los conceptos de excepción mayor y menor a una regla de clasificacion en la Sección 6.1.3. Para concluir, en la Sección 6.1.4, se propone una extensión bipolar del Método de Razonamiento Difuso presentado en 2.6.3.

6.1.1. RBC a nivel global en el marco de los SCBRDs

Como primer método para considerar un modelo de representación bipolar en SCBRDs, y siguiendo el enfoque a *Nivel Global* presentado en la Sección 5.1, nos centramos en la información contenida en las posibles regiones de confusión para la clasificación dada por un SCBRD básico.

Para esta exposición, se retomará la notación introducida en la Sección 2.6.3 al describir el Método de Razonamiento Difuso (MRD) de los SCBRD. En particular, se recuerda que, siguiendo a [16], en esa sección nos referimos a las puntuaciones intermedias que producen los SCBRD antes de la clasificación final como grados de consistencia (*soundness degrees*), denotándolos por $\pi_j(x)$. Estos grados de consistencia $\pi_j(x)$ proporcionan la evaluación final de los SCBRDs sobre la intensidad de la asociación entre la clase C_j y una instancia x a clasificar. En definitiva, el post-ajuste global aprovecha los grados de consistencia de la clasificación final $\pi_j(x)$, definidos en la Sección 2.6.3, de la instancia x a la clase C_j y lo manejamos en los términos dados en la Sección 5.1.

En estas regiones de confusión, para una instancia dada x, los grados de consistencia final $\pi_j(x)$ para algunas clases C_j pertenecientes al conjunto de clases S a menudo no son tan diferentes, y la clasificación final se basa, de alguna manera, en información imprecisa sobre la certeza de tomar dicha decisión. Por lo tanto, la aplicación del modelo bipolar propuesto podría hacer que el sistema clasifique estos casos inciertos de manera adecuada.

Se desarrolla a continuación la notación específica para este primer modelo SCBRD bipolar a *Nivel Global*. En primer lugar, para una clase dada $C_i \in S$ y una nueva instancia dada x, el grado de evidencia negativa $ev_i^-(x) = \pi_i^-(x)$ puede interpretarse como el grado de consistencia del patrón x hacia la clase abstracta definida por dC_i como aquel constructo que representa la unión de las clases de S diferente a C_i. Dicha evidencia negativa se puede calcular de la siguiente manera:

$$\pi_i^-(x) = \sum_{j \neq i} d_{ij} \pi_j^+(x) = \sum_{j=1}^{c} d_{ij}\ \pi_j^+(x) = D_i \pi^+(x) \tag{6.1}$$

donde $\pi^+(x) = (\pi_1^+(x), ..., \pi_c^+(x))^t$. Esta definición de evidencia negativa puede establecerse en notación matricial como sigue:

$$\pi^-(x) = D\pi^+(x) \tag{6.2}$$

Teniendo en cuenta los pares $(\pi_i^+(x), \pi_i^-(x))$ de grados de consistencia positivos y negativos para cada clase, el siguiente paso se dedica a obtener un solo grado ajustado para cada clase por medio de la aplicación de un operador de agregación, como los introducidos en la Sección 3.3.1, sobre los pares anteriores. Por lo tanto, podemos reformular la Definición (3.3.1) de la siguiente manera:

Definición 6.1.1. Sean $\pi_i^+(x)$, $\pi_i^-(x)$ los grados de consistencia positivo y negativo dados para una clase C_i y una instancia x. El grado de consistencia ajustado para la clase C_i se define como

$$\pi_i^{adj}(x) = \text{máx}\{0, \pi_i^+(x) - \pi_i^-(x)\}. \tag{6.3}$$

A partir de estos grados de consistencia ajustados, se realiza una asignación de clase mediante la regla del máximo, es decir, el elemento x se asignará a la clase C_h con un grado de consistencia ajustado máximo $\pi_h^{adj}(x)$.

6.1.2. Ajuste bipolar de los grados de certeza de las reglas

Como se introdujo anteriormente, el modelo de representación bipolar local propuesto para SCBRD se entiende como un proceso de ajuste de la certeza de la regla, en el sentido de modificar la fuerza de la asociación positiva entre cada regla en la base de reglas (BR) y las diferentes clases en S. Por este motivo, este enfoque local se debe aplicar tras la etapa de aprendizaje de la BR del SCBRD (marcado en azul en la Figura 6.1), en lugar de al final del proceso de MRD (en verde en la Figura 6.1) como lo hace el enfoque global recién expuesto.

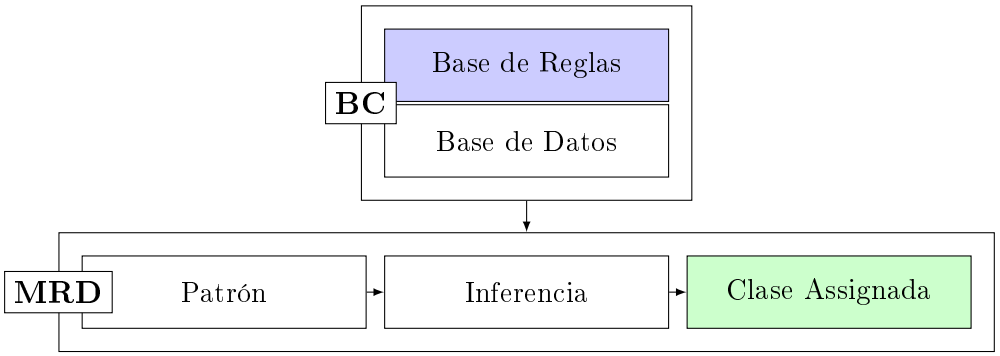


Figura 6.1: Diagrama de flujo en dos etapas de un SCBRD. Primera etapa (BC + BD): aprendizaje y creación la BR y la BC. Segunda etapa (MRD): asignación final de clase para nuevos elementos mediante un proceso de inferencia.

Sin pérdida de generalidad, es posible asumir un SCBRD base con una BR formada por reglas de tipo III, como se muestra en la Sección 2.6.2. Como

las reglas de tipo I y tipo II pueden considerarse como casos particulares de reglas de tipo III, la siguiente formulación dada para reglas de tipo III puede ser aplicar a cualquier BR formada por reglas de tipo I o tipo II.

Así pues, representemos con $r_i^{q,-}$ el grado de confianza de la regla $R_i^{q,-}$: $A^q \Rightarrow dC_i$. Este grado es obviamente interpretado como evidencia negativa, que "empuja" la clasificación en contra de la asociación del antecedente A^q con la clase C_i, ya que proporciona evidencia de que dicho antecedente puede estar asociado con clases disímiles a C_i.

Una forma natural de obtener este grado de confianza negativo $r_i^{q,-}$ es considerar que surge de la aplicación de la estructura de disimilitud a los grados de confianza positivos $r_j^{q,+}$, $j \neq i$. Siguiendo el marco general presentado en el Capítulo 3, consideramos la definición dada en la siguiente expresión,

$$r_i^{q,-} = \sum_{j \neq i} d_{ij}\ r_j^{q,+} = \sum_{j=1}^{c} d_{ij}\ r_j^{q,+} = D_i r^{q,+}, \tag{6.4}$$

con $r^{q,+} = (r_1^{q,+}, ..., r_c^{q,+})$ y donde D_i denota la fila i de la matriz de disimilitud considerada D. De nuevo, esta definición de grados de certeza negativa se puede establecer en notación matricial de la siguiente manera:

$$r^{q,-} = D r^{q,+}. \tag{6.5}$$

Teniendo en cuenta el par $(r_i^{q,+}, r_i^{q,-})$ de grados de certeza de la regla positiva y negativa para cada clase y cada regla R^q en la BR, el siguiente paso está dedicado a la obtención de un solo grado ajustado para cada clase y regla mediante la aplicación de un operador de agregación en el par anterior. Como se muestra en la Sección 4.1.2, el operador de agregación seleccionado se basa en la t-norma de Lukasiewicz, por lo que podemos re formular la Definición 3.3.1 de la siguiente manera:

Definición 6.1.2. Sean $r_i^{q,+}$, $r_i^{q,-}$ los grados de certeza o pesos de las reglas positivo y negativo, respectivamente, para la clase C_i dados por la regla R^q. El grado de certeza aditivo ajustado para la clase C_i en la regla R^q se define como

$$r_i^{q,adj} = \text{máx}\{0, r_i^{q,+} - r_i^{q,-}\}. \tag{6.6}$$

Observe que estos pesos de reglas ajustados se pueden usar como pesos de reglas habituales, en el sentido de que proporcionan un grado de certeza unidimensional para cada clase que coincide con la entrada necesaria para el marco de MRD estándar expuesto en la Sección 2.6.3.

6.1.3. Distinguiendo entre excepciones menores y mayores en una regla de clasificación

Ahora exploremos desde una perspectiva general la interpretación de la introducción de relaciones de disimilitud entre las diferentes clases en el marco de la clasificación basada en reglas siguiendo el enfoque local recién expuesto. La idea principal aquí es que este tipo de relación puede permitir considerar algunas clases como más excepcionales que otras en relación con el consecuente de una regla de clasificación dada.

Entonces, digamos que U_X denota el espacio de entrada dado por el producto cartesiano de los rangos U_{X_i} de las variables independientes $X_1, ..., X_n$, que es $U_X = U_{X_1} \times ... \times U_{X_n}$ y sea $S = \{C_1, ..., C_c\}$ el conjunto de c clases consideradas en nuestro problema de clasificación. Además, supongamos que $x = (x_1, ..., x_n) \in U_X$ representa una instancia a la que se le debe asignar una clase en S. Y finalmente sea $A \subset U_X$ una región del espacio de entrada, que por lo tanto puede constituir la premisa o el antecedente de una regla de clasificación en la forma $R_j : A \Rightarrow C_j$, $j \in 1, ..., c$, que expresa la relación "si $x \in A$, entonces x debe asignarse a la clase C_j".

Así pues, en estas condiciones y, de acuerdo con [20], es importante subrayar que una regla de la forma $R_j : A \Rightarrow C_j$ produce una 3-partición borrosa del espacio de datos $U = U_X \times S$ dada por:

- El conjunto de ejemplos de la regla, $A \wedge C_j$.
- Su conjunto de contraejemplos, $A \wedge C_j^c$.
- El conjunto de patrones irrelevantes, A^c.

Por lo tanto, la calidad de tal regla se puede evaluar por medio de dos cantidades independientes, las proporciones de los ejemplos positivos y negativos, respectivamente $P(A \cap C_j)$ y $P(A \cap C_j^c)$. Esta concepción de una regla es equivalente a la que generalmente se asume en la minería de datos y el aprendizaje automático, en el que la validez de una regla se evalúa en términos de dos cantidades independientes, el soporte $P(A)$ y la confianza $P(C_j|A)$, ya que

$$P(A) = P(A \cap C_j) + P(A \cap C_j^c)$$

y

$$P(C_j|A) = \frac{P(A \cap C_j)}{P(A \cap C_j) + P(A \cap C_j^c)}$$

Sin embargo, obsérvese que al comparar varias sub-reglas $R_j : A \Rightarrow C_j$, $j = 1, ..., c$, es decir, reglas que tienen la misma región de antecedente pero diferentes clases consecuentes (lo cual es bastante habitual en el contexto de la clasificación basada en reglas), las proporciones mencionadas de ejemplos

positivos y negativos simplemente proporcionan la misma información. Esto sucede porque $P(A \cap C_j)$ es solo el complemento de $P(A \cap C_j^c)$ con respecto a $P(A)$, pues $P(A \cap C_j) = P(A) - P(A \cap C_j^c)$ o, equivalentemente,

$$P(C_j|A) = \frac{P(A \cap C_j)}{P(A)} = 1 - \frac{P(A \cap C_j^c)}{P(A)} = 1 - P(C_j^c|A)$$

En consecuencia, dado un antecedente A y una regla $R_j : A \Rightarrow C_j$, todos los ejemplos de aprendizaje contenidos en la celda $A \times S$ se consideran positivos o negativos, es decir, todas las instancias de entrenamiento que no son ejemplos favorables a la regla son consideradas como contrajejemplos, contrarios a tal regla. Por lo tanto, se asume implícitamente que cada clase C_j es igual y totalmente opuesta a las otras clases en S.

No obstante, esta suposición se puede relajar al otorgar al conjunto de clases S una relación de disimilitud. Es decir, si se considera que una clase C_i en S es más disimil a una clase dada C_j que las otras clases en S, entonces los contraejemplos en $A \wedge C_i$ constituirán excepciones más significativas a la regla $R_j : A \Rightarrow C_j$ que otros contraejemplos en $A \wedge (C_i \vee C_j)^c$. En otras palabras, la introducción de una estructura de disimilitud en el conjunto de clases hace posible separar los contraejemplos de una regla $R_j : A \Rightarrow C_j$ en excepciones menores y significativas a dicha regla. Esto permite distinguir la situación en la que una regla tiene una cantidad representativa de excepciones significativas de aquella otra situación en la que solo tiene excepciones menores. La confianza de la regla debe ser menor en la primera situación que en la segunda.

En este sentido, la idea central en la propuesta de [68, 70] era que la introducción de algunas suposiciones de disimilitud en el conjunto de clases permitía modelar algunos requisitos decisionales relevantes del contexto de la aplicación (que en esos trabajos era la gestión de desastres, con las clases representando diferentes niveles de magnitud de las consecuencias del desastre), ya que la presencia de contraejemplos significativos de una regla $R_j : A \Rightarrow C_j$ cerca de algunos ejemplos constituía un hecho importante a tener en cuenta para evaluar la validez de dicha regla. Sin embargo, en este documento, como se mencionó anteriormente, seguimos el enfoque en [81], que consiste en buscar los supuestos de disimilitud más convenientes en el conjunto de clases para mejorar la capacidad predictiva del clasificador. Es decir, buscamos la separación más adecuada de contraejemplos entre excepciones menores y significativas de cara a optimizar el rendimiento del clasificador.

6.1.4. Extensión bipolar del MRD de los SCBRDs

Se propone ahora una extensión del método de razonamiento difuso (MRD) general [16] expuesto en la Sección 2.6.3 para permitir el manejo de grados

de certeza positivos y negativos de reglas al clasificar una nueva instancia x. Aunque la versión propuesta del enfoque local presentado en este trabajo agrega los pesos de las reglas positiva y negativa en un solo peso ajustado, tales grados de certeza bipolar pueden explotarse de diferentes maneras, particularmente sin tal paso de agregación. Por lo tanto, la extensión del MRD propuesta en esta sección proporciona un marco en el que son posibles estrategias de explotación más generales que el esquema de agregación utilizado en este trabajo. En particular, los enfoques locales y globales expuestos (con su correspondiente agregación de evidencia positiva y negativa) pueden considerarse casos especiales de la extensión del MRD estándar propuesta a continuación.

Esta extensión del MRD para SCBRDs bipolares lleva a cabo los siguientes pasos cuando se considera una instancia $x = (x_1, ..., x_n)$ a clasificar en una de las c clases disponibles, asumiendo que se tiene una BR con N_R reglas de tipo III:

- *Grado de emparejamiento* del patrón x y la parte antecedente de cada regla R^q en BR,

$$\sigma_{R^q}(x) = T(\mu_{A_1^q}(x_1), ..., \mu_{A_n^q}(x_n)),\ q = 1, ..., N_R$$

  donde T denota un operador de conjunción como una t-norma o una función *overlap*.

- *Grados de asociación positiva y negativa* del patrón x con la clase C_j de acuerdo con cada regla R^q, obtenida al agregar por separado (generalmente a través del operador del producto) el grado de emparejamiento anterior del patrón y los grados de certeza o pesos de reglas positivo y negativo para la clase j en el consecuente de la regla:

$$b_j^{q,+}(x) = h^+(\sigma_{R^q}(x), r_j^{q,+}),\ j = 1, .., c,\ q = 1, ..., N_R$$
$$b_j^{q,-}(x) = h^-(\sigma_{R^q}(x), r_j^{q,-}),\ j = 1, .., c,\ q = 1, ..., N_R$$

- *Función de ponderación* en la forma $g^+, g^- : [0,1] \times [0,1]$ que potencia asociaciones altas y penaliza aquellas de menor valor.

$$w_j^{q,+}(x) = g^+(b_j^{q,+}(x))$$
$$w_j^{q,-}(x) = g^-(b_j^{q,-}(x))$$

- *Grado de consistencia o consistencia del patrón de clasificación* para todas las clases, calculado por medio de la aplicación de funciones de agregación f^+, f^- que combinen, para cada clase C_j, los grados de

asociación positivos ponderados $w_j^q(x)$ de todas las reglas calculadas en etapas previas.

$$\pi_j^+(x) = f^+((w_j^{q,+}(x), w_j^{q,-}(x)), q = 1, ..., N_R),\ j = 1, ..., c$$
$$\pi_j^-(x) = f^-((w_j^{q,+}(x), w_j^{q,-}(x)), q = 1, ..., N_R),\ j = 1, ..., c$$

- *Clasificación nítida final*, producida a través de un proceso de *defuzzificación* que transforma el grado de consistencia de todas las clases en una única asignación, por medio de la aplicación de una función de decisión F que actúa sobre una agregación ϕ de los pares de grados de consistencia para cada clase. La función de decisión F se puede elegir para aplicar simplemente la regla del máximo sobre estas agregaciones de los pares de grados de consistencia positivos y negativos, es decir,

$$C_h = F(\phi(\pi_1^+(x), \pi_1^-(x)), .., \phi(\pi_c^+(x), \pi_c^-(x)))$$

tal que

$$\phi(\pi_h^+(x), \pi_h^-(x)) = \max_{j=1,..,c} \phi(\pi_j^+(x), \pi_j^-(x))$$

Es conveniente señalar que tanto los métodos globales como los locales expuestos anteriormente, pueden interpretarse como casos particulares de este MRD extendido.

En el caso del método global, los grados de consistencia negativos π^- pueden considerarse como obtenidos a través de una función de agregación f^- que introduce la estructura de disimilitud (siguiendo la Ecuación (6.1)) en la etapa de agregación del MRD. Y se puede considerar que la agregación tipo Lukasiewicz de grados de consistencia positiva y negativa en la Ecuación (6.3) se realiza a través del operador ϕ en el último paso de decisión.

Con respecto al método local, se puede considerar que la agregación tipo Lukasiewicz de grados de certeza positivos y negativos en la Ecuación (6.6) también se puede lograr a través de la función f^+ en el paso de agregación del MRD extendido, ya que no se consideran grados de consistencia negativos en este caso.

Por lo tanto, de la sinergia entre estas dos estrategias basadas en humanos para el manejo de la información (SCBRDs y RBC), surge un nuevo marco bipolar para la inferencia borrosa.

6.2. Estudio experimental

Con el objetivo de evaluar el comportamiento de la propuesta descrita anteriormente, se lleva a cabo un completo estudio experimental. En primer lugar, en la Sección 6.2.1 se detalla la configuración experimental escogida.

La Sección 6.2.2 presenta un caso de estudio teórico que arroja luz sobre el alcance de la actuación del nuevo paradigma bipolar a nivel de regla para permitir el manejo de pesos o grados de certeza de las reglas de carácter bipolar. Finalmente en la Sección 6.2.3 se muestran los resultados de un completo experimento con datos reales llevado a cabo para estudiar el rendimiento de esta nueva aproximación bipolar en el marco de los SCBRDs.

6.2.1. Configuración experimental

Con el objetivo de evaluar el comportamiento del nuevo método local recientemente propuesto, se lleva a cabo un experimento computacional en el que se considera el clasificador de Chi como SCBRD base, por lo que en primer lugar se dan los detalles de este algoritmo. A continuación, se proporciona información sobre el análisis experimental, como los conjuntos de datos considerados y las pruebas estadísticas aplicadas.

6.2.1.1. Detalles del SCBRD

En efecto, se hace uso en este punto del paquete *frbs()* implementado en R [74], que permite al usuario ajustar distintos SCBRDs a cualquier conjunto de datos apropiado. La elección de esta implementación en concreto, se relaciona con la necesidad de introducir, en el código fuente de la misma, las modificaciones necesarias para la correcta aplicación del paradigma de representación bipolar. Concretamente, se retoca el código principal con el objetivo de obtener los grados de evidencia en ambos niveles *global* y *local*.

Se decide, por tanto, trabajar sobre el clasificador basado en partición del espacio de Chi [14], un algoritmo ampliamente utilizado con una estructura de aprendizaje adecuada para evaluar el comportamiento de la inferencia bipolar propuesta. De acuerdo con [74], el método de Chi extiende aquel de Wang y Mendel [85] para abordar los problemas de clasificación. En primer lugar, se desarrolla un esquema de partición de espacio basado en el número de etiquetas seleccionadas. A continuación, se crean reglas difusas para cada ejemplo de entrenamiento, lo que lleva a considerar de manera implícita reglas con múltiples clases consecuentes, es decir, reglas de tipo III. De hecho, es bastante similar a la técnica de Wang y Mendel. Para más detalles sobre este método y su implementación en R Software, vea [14] y [74].

El clasificador de Chi considerado en este estudio tiene la siguiente configuración paramétrica:

- Número de etiquetas lingüísticas = 3
- Tipo de función de pertenencia = *Triangular*
- Tipo de función de implicación, h = *Producto*

- Tipo de T-norma, $T = Min$
- Tipo de T-conorma, $f = Max$

Estableciendo una conexión con el método de razonamiento difuso (MRD) presentado en la Sección 2.6.3, se tienen tres funciones de pertenencia triangulares (con etiquetas *small*, *medium* and *large*) para cada característica, se considera la t-norma mínimo para calcular el grado de emparejamiento (esto es, el MRD utiliza $T =$ mín), la función de implicación producto para calcular los grados de asociación (esto es, $h = prod$), no se considera función de ponderación ($g = identity$), la agregación se realiza por regla ganadora ($f = max$) y la asignación final de clase mediante la regla del máximo.

En cuanto a la estructura de disimilitud, se utiliza el esquema de aprendizaje propuesto en la Sección 3.5, por medio de un AG implementado en [84] con los siguientes parámetros:

- Tamaño de la población: 50 individuos
- Número de iteraciones: 20
- *Mutation chance* (la posibilidad de que un gen en el cromosoma mute): 0.01
- *Elitism* (el número de cromosomas que se mantienen en la próxima generación): aproximadamente el 20 % del tamaño de la población

6.2.1.2. Conjuntos de datos

En este caso, se seleccionan 30 conjuntos de datos del repositorio KEEL [75]. En la misma línea que en experimentos anteriores se hace uso de un esquema de validación cruzada *5-folds*. La Tabla 6.1 resume la información sobre los seleccionado conjuntos de datos, mostrando en primer lugar las categorías escogidas como *positiva* C_2 y *negativa* C_1, el número de ejemplos (#Ex.), el número de atributos (#Atts.) y su tipo (Real/Integer/Nominal). Para convertir un conjunto de datos multiclase en un conjunto de datos de dos clases (C_1, C_2), tomamos como clase C_2 la clase original más cercana a 20 % de instancias, y como clase C_1 la unión de las clases restantes.

6.2.1.3. Análisis estadístico

En este análisis, se utilizan algunas técnicas de validación de hipótesis para dar soporte estadístico al análisis de los resultados. Se consideran pruebas no paramétricas de Friedman y Wilcoxon ya descritas en la Sección 4.2.1.

Una completa descripción de estas pruebas junto con muchas consideraciones y recomendaciones, e incluso el software utilizado para ejecutar este análisis se encuentra disponible en el sitio web `sci2s.ugr.es/sicidim`.

Id.	Data-set	Clase C_2 (+)	#Ex.	#Atts.	(R/I/N)
Aba	Abalone	9	4174	8	(7/1/0)
App	Appendicitis	0	106	7	(7/0/0)
Aus	Australian	0	690	14	(3/5/6)
Bal	Balance	L	625	4	(4/0/0)
Ban	Banana	-1	5300	2	(2/0/0)
Bup	Bupa	1	345	6	(1/5/0)
Car	Car	acc	159	25	(15/0/10)
Con	Contraceptive	2	1473	9	(6/0/3)
Eco	Ecoli	im	336	7	(7/0/0)
Gla	Glass	1	214	9	(9/0/0)
Hay	Hayes-Roth	3	160	4	(0/4/0)
Hea	Heart	1	270	13	(1/12/0)
Iri	Iris	Setosa	150	4	(4/0/0)
Led	Led7digit	3	500	7	(7/0/0)
Lin	Lymphography	malign_lymph	148	18	(3/0/15)
Mag	Magic	g	19020	10	(10/0/0)
Nty	Newthyroid	2	215	5	(4/1/0)
Pag	Page-blocks	2	5472	10	(4/6/0)
Pen	Penbased	0	10992	16	(0/16/0)
Pho	Phoneme	0	5404	5	(5/0/0)
Pim	Pima	tested_negative	768	8	(8/0/0)
Sah	Saheart	0	462	9	(5/3/1)
Tit	Titanic	-1.0	2201	3	(3/0/0)
Veh	Vehicle1	positive	846	18	(0/18/0)
Vow	Vowel	0	990	13	(10/3/0)
Win	Wine	1	178	13	(13/0/0)
Wqr	Winequality-red	7	1599	11	(11/0/0)
Wqw	Winequality-white	7	4898	11	(11/0/0)
Wis	Wisconsin	2	699	9	(0/9/0)
Yea	Yeast	MIT	1484	8	(8/0/0)

Tabla 6.1: Descripción de los conjuntos de datos empleados en la propuesta de marco bipolar en el contexto de los SCBRDs.

6.2.2. Caso de estudio teórico

Para ilustrar el comportamiento del enfoque bipolar propuesto en el contexto de los SCBRDs, hemos seleccionado algunos ejemplos del estudio experimental que se acaba de presentar, usando el método de Chi [14] como base SCBRD. Particularmente, primero mostramos cómo se modifican los grados de consistencia finales de algunas instancias de prueba después de aplicar el ajuste global descrito en la Sección 6.1.1. Luego, las mismas instancias de prueba se utilizan para ilustrar cómo el ajuste local (Sección 6.1.2) varía los grados de certeza de las reglas difusas aprendidas por el clasificador. En ambos casos, el ajuste bipolar realizado después del proceso de aprendizaje de disimilitud conduce a modificar la asignación de clase inicial del clasificador,

lo que permite una mejora de su rendimiento.

Para este caso de estudio, se considera el conjunto de datos *banana*, que plantea un problema bidimensional, esto es con dos variables explicativas, de clasificación binaria (consulte la Tabla 6.1 para obtener más detalles). Dado que se trata de un conjunto de datos bidimensionales, es posible representar las áreas y fronteras de clasificación dados por la BR a través de una imagen 2D, algo que puede servir para ilustrar algunos aspectos del procedimiento propuesto. Como se expuso anteriormente, el algoritmo Chi de referencia se aplica con la configuración detallada en la Sección 6.2.1.1.

La BR dada por el método de Chi tras el ajuste a la muestra de entrenamiento del conjunto *banana* contiene las siguientes 8 reglas de tipo III:
R^1: Si X_1 is *medium* and X_2 is *medium* entonces $(.89, .98)$
R^2: Si X_1 is *small* and X_2 is *medium* entonces $(.89, .67)$
R^3: Si X_1 is *medium* and X_2 is *large* entonces $(.84, .64)$
R^4: Si X_1 is *large* and X_2 is *medium* entonces $(.74, .83)$
R^5: Si X_1 is *large* and X_2 is *small* entonces $(.80, .61)$
R^6: Si X_1 is *medium* and X_2 is *small* entonces $(.76, .67)$
R^7: Si X_1 is *small* and X_2 is *small* entonces $(0, .71)$
R^8: Si X_1 is *large* and X_2 is *large* entonces $(0, .50)$

Como podemos ver, esta base de reglas tiene 6 reglas con doble consecuente y solo 2 reglas de consecuente único. Por lo tanto, hay varios subespacios superpuestos en los datos, en los que la región cubierta por una regla contiene instancias pertenecientes a diferentes clases. Como el MRD de Chi utiliza la agregación por regla ganadora de los grados de asociación, la asignación final de clase de una nueva instancia x en estas regiones está fuertemente influenciada por los grados de certeza de las reglas que mejor se ajustan a estas regiones. En otras palabras, los grados de consistencia de la instancia x, que determinan su asignación de clase, serán bastante similares a los grados de certeza de la regla con el antecedente que mejor coincida con x. En este contexto, el enfoque local propuesto modifica indirectamente estos grados de consistencia mediante el ajuste de los grados de certeza de las reglas de forma similar a como se ajustan los pesos de las reglas en [46], mientras que el enfoque global actúa directamente sobre los grados de consistencia.

Una vez señalada la importancia de los pesos de las reglas en el proceso de inferencia, los siguientes ejemplos ilustran la clasificación de dos instancias en la muestra de prueba del conjunto *banana* utilizando los dos métodos propuestos de ajuste bipolar basados en estructuras de disimilitud.

Por lo tanto, denotemos por x_1 y x_{251} las instancias 1 y 251 del conjunto de prueba de *banana*, con valores de características de entrada $x_1 = (1.83, .45)$ y $x_{251} = (1.67, -.98)$, siendo $[-3.09, 2.81]$ y $[-2.39, 3.19]$ los rangos de características X_1 y X_2, respectivamente. La Tabla 6.2 muestra el proceso de fuzzificación de ambas características, así como los grados de

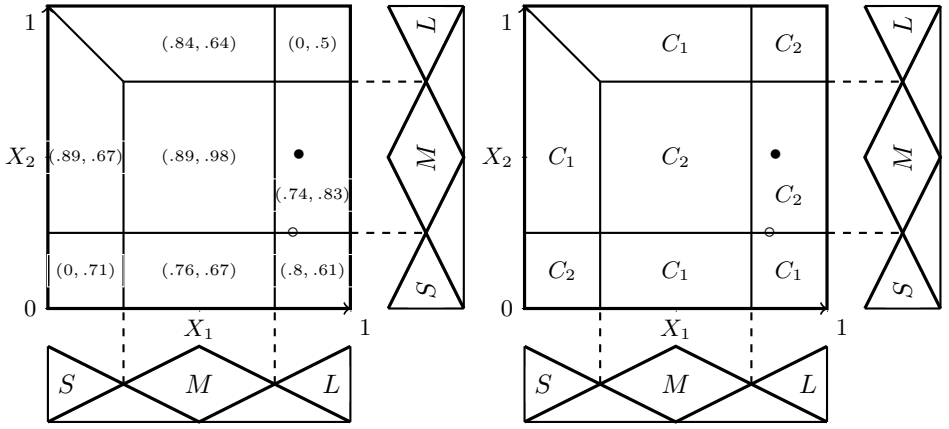


Figura 6.2: Regiones del espacio de características asociadas a los grados originales de las 8 reglas en la BR proporcionada por el clasificador de Chi en el conjunto *banana*. Las instancias x_1 y x_{251} se muestran respectivamente con un punto negro y un círculo.

pertenencia de estos valores en los tres conjuntos difusos utilizados para describir lingüísticamente los valores de ambas características (s significa *small*, m para *medium* y l para *large*). Téngase en cuenta también que la verdadera clase de ambas instancias es C_1.

i	X_1	X_2	$X_1^{Label}(s,m,l)$	$X_2^{Label}(s,m,l)$	TrueClass
1	1.83	.45	(0, .33, .67)	(0, .98, .02)	C_1
251	1.67	-.98	(0, .39, .61)	(.49, .51, 0)	C_1

Tabla 6.2: Proceso de fuzzification de las instancias x_1 y x_{251}

Para poder asignar cada elemento a una determinada clase a través de su mecanismo de inferencia, el primer paso que realiza el SCBRD es una *fuzzificación* de los valores de entrada, para transformar la información contenida en los datos en grados de verificación de los correspondientes términos lingüísticos.

Ea necesario resaltar que ambas instancias son bastante similares en términos de la característica X_1, y están asociadas predominantemente con la etiqueta *large*. Con respecto a X_2, la instancia x_1 presenta claramente un valor de *medium*, mientras que x_{251} se encuentra casi en el límite entre *small* y *medium*. Como se muestra en la Figura 6.2, esto hace de x_{251} una instancia límite, situada en la frontera de decisión del clasificador de Chi entre las clases C_1 y C_2.

Como se muestra en la Tabla 6.3, que ilustra los detalles de la aplicación del MRD, los grados de consistencia obtenidos para ambas instancias son, respectivamente $\pi^+(x_1) = (.49, .55)$ y $\pi^+(x_{251}) = (.39, .42)$. Como resultado,

ambas instancias quedan mal clasificadas en la clase C_2.

Regla	Gr. Compat.	Gr. Emp.	Gr. Asoc.
R^q	$\mu^q_{A_i}(x)$	$\sigma_{R^q}(x)$	$b^q(x)$
R^1	(.33, .98)	.33	(.29, .32)
R^2	(0, .98)	0	(0, 0)
R^3	(.33, .02)	.02	(.015, .011)
R^4	(.67, .98)	.67	(.49, .55)
R^5	(.67, 0)	0	(0, 0)
R^6	(.33, 0)	0	(0, 0)
R^7	(0, 0)	0	(0, 0)
R^8	(.67, .02)	.02	(0, .01)
Grado de consistencia final			(.49, **.55**)
Asignación de clase			Clase C_2

Regla	Gr. Compat.	Gr. Emp.	Gr. Asoc.
R^q	$\mu^q_{A_i}(x)$	$\sigma_{R^q}(x)$	$b^q(x)$
R^1	(.39, .51)	.39	(.34, .38)
R^2	(0, .51)	0	(0, 0)
R^3	(.39, 0)	0	(0, 0)
R^4	(.61, .51)	.51	(.38, .42)
R^5	(.61, .49)	.49	(.39, .30)
R^6	(.39, .49)	.39	(.29, .26)
R^7	(0, .49)	0	(0, 0)
R^8	(.61, 0)	0	(0, 0)
Grado de consistencia final			(.39, **.42**)
Asignación de clase			Clase C_2

Tabla 6.3: Proceso completo de inferencia. Instancia x_1 (arriba) y x_{251} (abajo)

Enfoque Global. La siguiente matriz de disimilitud se obtuvo después de realizar el aprendizaje genético de la estructura de disimilitud en la muestra de entrenamiento del conjunto de datos *banana*, utilizando el enfoque global expuesto en la Sección 6.1.1:

$$D_{Global} = \begin{bmatrix} 0 & .68 \\ .82 & 0 \end{bmatrix}$$

Así, la matriz anterior se aplicará sobre los grados de consistencia (positivos) $\pi^+(x) = \pi(x) = (\pi_1(x), \pi_2(x))$ obtenidos al final del proceso de MRD. La aplicación de tal estructura de disimilitud a estas evidencias positivas proporciona los siguientes grados de consistencia negativa para cada clase:

$$\pi^{-}(x_1) = D_{Global}\pi^{+}(x_1) = \begin{bmatrix} 0 & .68 \\ .82 & 0 \end{bmatrix} \begin{bmatrix} .49 \\ .55 \end{bmatrix} = \begin{bmatrix} .38 \\ .41 \end{bmatrix}$$

$$\pi^{-}(x_{251}) = D_{Global}\pi^{+}(x_{251}) = \begin{bmatrix} 0 & .68 \\ .82 & 0 \end{bmatrix} \begin{bmatrix} .39 \\ .42 \end{bmatrix} = \begin{bmatrix} .28 \\ .32 \end{bmatrix}$$

Como se describe en la Sección 6.1.1, estos grados de consistencia negativos se pueden usar para corregir los positivos atendiendo a las relaciones de disimilitud aprendidas entre las clases. Esta corrección se realiza a través del esquema de agregación aditiva, lo que proporciona los siguientes grados de consistencia ajustados:

$$\pi^{adj}(x_1) = \text{máx}\{0, \pi^{+}(x_1) - \pi^{-}(x_1)\} = \text{máx}\{\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} .49 \\ .55 \end{bmatrix} - \begin{bmatrix} .38 \\ .41 \end{bmatrix}\} = \begin{bmatrix} .11 \\ \mathbf{.14} \end{bmatrix}$$

$$\pi^{adj}(x_{251}) = \text{máx}\{0, \pi^{+}(x_{251}) - \pi^{-}(x_{251})\} = \text{máx}\{\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} .39 \\ .42 \end{bmatrix} - \begin{bmatrix} .28 \\ .32 \end{bmatrix}\} = \begin{bmatrix} \mathbf{.11} \\ .10 \end{bmatrix}$$

Por lo tanto, después del ajuste global del clasificador Chi, la instancia x_1 se asigna a la clase C_2, y x_{251} a la clase C_1. Es importante señalar que, para cada matriz D tentativa en la búsqueda genética, los grados de consistencia de todos las instancias de entrenamiento se ajustan como se describe en este ejemplo, y la matriz D_{Global} presentada anteriormente es la matriz D que produjo la mayor mejora del índice kappa sobre los datos de entrenamiento.

Enfoque Local. Ilustremos ahora el procedimiento de ajuste de reglas expuesto anteriormente expuesto y su efecto en la clasificación de las instancias de prueba x_1 y x_{251} del conjunto *banana*. En este caso, la estructura de disimilitud aprendida a través de la búsqueda de AG está dada por la matriz

$$D_{Local} = \begin{bmatrix} 0 & .68 \\ .97 & 0 \end{bmatrix}$$

A diferencia del enfoque global, esta matriz debe aplicarse sobre los grados de certeza (positivos) $r_j^{+} = r_j,\ j = 1, 2$ de cada una de las 8 reglas proporcionadas por el método de Chi dadas arriba, con el fin de producir los grados de certeza negativos para cada clase. Este proceso se detalla a continuación para la primera regla R^1 en la BR:

$$r_{R_1}^{-} = D_{Local} r_{R_1}^{+} = \begin{bmatrix} 0 & .68 \\ .97 & 0 \end{bmatrix} \begin{bmatrix} .89 \\ .98 \end{bmatrix} = \begin{bmatrix} .67 \\ .86 \end{bmatrix}$$

Siguiendo el método expuesto en la Sección 6.1.2, estos grados de certeza negativa son capaces de corregir los grados de certeza de la regla R^1 atendiendo a las relaciones de disimilitud aprendidas entre las clases. Esta corrección se realiza a través del esquema de agregación aditiva, lo que lleva a los siguientes grados de certeza ajustados para la regla R^1:

$$r_{R^1}^{adj} = \text{máx}\{0, r_{R_1}^{+} - r_{R_1}^{-}\} = \text{máx}\{\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} .89 \\ .98 \end{bmatrix} - \begin{bmatrix} .67 \\ .86 \end{bmatrix}\} = \begin{bmatrix} .21 \\ .12 \end{bmatrix}$$

La Tabla 6.4 resume el resultado de este proceso para cada una de las 8 reglas en la BR. Por lo tanto, el enfoque local propuesto se centra en ajustar los grados de certeza de la regla, penalizando de alguna manera a aquellos que coexisten (en el consecuente de la regla de tipo III) con grados de certeza positivos (es decir, mayores que cero) hacia clases disimiles. De esta manera, este proceso de ajuste de reglas puede cambiar las fronteras de decisión dadas por la BR inicial, como se puede ver en la Figura 6.3 comparando los grados de certeza iniciales de las reglas y los obtenidos después del ajuste.

R^i	r^+	r^-	r^{adj}
R^1	(.89, .98)	(.67, .86)	(.21, .12)
R^2	(.89, .67)	(.46 .86)	(.43, 0)
R^3	(.84, .64)	(.44 .82)	(.40, 0)
R^4	(.74, .83)	(.57 .72)	(.17, .10)
R^5	(.80, .61)	(.42 .78)	(.38, 0)
R^6	(.76, .67)	(.46 .74)	(.30, 0)
R^7	(0, .71)	(.49, 0)	(0, .71)
R^8	(0, .50)	(.34, 0)	(0, .50)

Tabla 6.4: Ajuste bipolar de los grados de certeza o pesos de las reglas.

Lo que queda por hacer en este punto, es aplicar el MRD con la BR ajustada mediante el paradigma bipolar a nivel local a las instancias de prueba x_1 y x_{251} para asignarles una clase. La información de todo el proceso del MRD se muestra en la Tabla 6.5, incluido el cálculo de los correspondientes grados de emparejamiento, asociación y consistencia, así como la asignación final de la clase, para el clasificador ajustado localmente. Observe que, como resultado del ajuste local de los grados de certeza de la regla, ambas instancias de prueba se clasifican ahora correctamente en la clase C_1.

Como antes, es importante remarcar que la matriz D_{Local} se aprendió usando los procesos de ajuste e inferencia ilustrados sobre la muestra de entrenamiento. Es decir, para cada matriz D probada en la búsqueda del AG, los grados de certeza de las 8 reglas se ajustaron como se acaba de describir. Después del ajuste, todos los ejemplos de prueba se clasifican aplicando el MRD en la BR ajustada, y la matriz D_{Local} presentada anteriormente es solo la matriz D que produjo la mayor mejora en el índice kappa en los datos de entrenamiento.

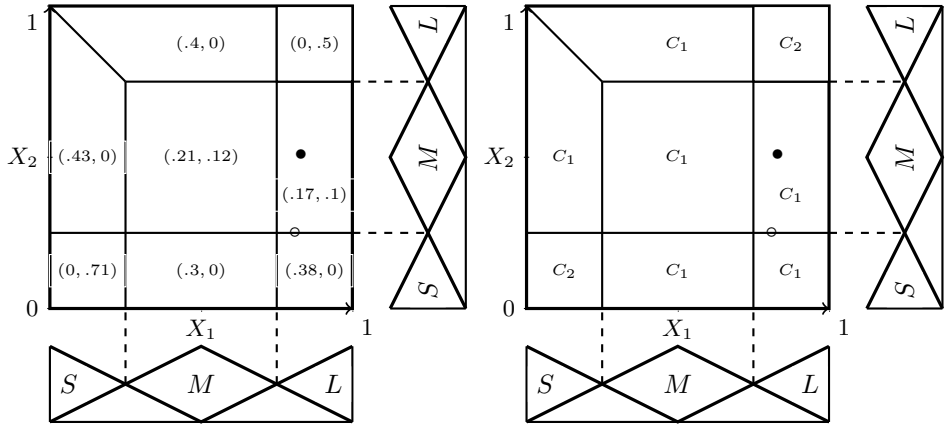


Figura 6.3: Regiones del espacio de características asociadas a los grados de certeza ajustados por bipolaridad de las 8 reglas en la BR proporcionada por el clasificador de Chi en el conjunto *banana*. Las instancias x_1 y x_{251} se muestran respectivamente con un punto negro y un círculo.

Como conclusiones de este estudio teórico, se considera el resultado de ajustar un clasificador de Chi a la muestra de entrenamiento del conjunto de datos *banana* y su BR de 8 reglas de tipo III. Luego nos enfocamos en la aplicación del modelo ajustado a dos instancias de prueba, x_1 y x_{251}, ambas clasificadas erróneamente por el clasificador de Chi. En este punto, se aplican los dos enfoques de ajuste bipolar propuestos, lo que proporciona dos estructuras de disimilitud diferentes que permiten obtener las evidencias negativas correspondientes. Estas evidencias negativas se utilizaron a su vez para corregir los resultados del clasificador de base atendiendo a las estructuras de disimilitud aprendidas: los grados de consistencia en el caso del enfoque global y los grados de certeza de las 8 reglas en el caso del enfoque local. Los dos clasificadores ajustados se utilizaron para predecir las dos instancias de prueba, y los resultados se resumen en la Tabla 6.6. En particular, el enfoque global permitió corregir la clasificación errónea de la instancia x_{251}, mientras que el enfoque local también permitió clasificar correctamente la instancia x_1.

i	Clase real	π^{Chi}	C^{Chi}	$\pi^{Chi_{glo}}$	$C^{Chi_{glo}}$	$\mu^{Chi_{loc}}$	$C^{Chi_{loc}}$
1	C_1	(.49, .55)	C_2	(.11, .14)	C_2	(.12, .07)	C_1
251	C_1	(.39, .42)	C_2	(.11, .10)	C_1	(.19, .05)	C_1

Tabla 6.6: Clasificaciones original y ajustadas para las instancias x_1 y x_{251}.

Observemos en este punto las diferentes filosofías de los dos enfoques propuestos de ajuste bipolar. Si bien el enfoque global se aplica sobre los grados de consistencia finales de la muestra de entrenamiento, el ajuste local se cen-

Regla	Gr. Compat.	Gr. Emp.	Nuevo Gr. Asoc.
R^q	$\mu^q_{A_i}(x)$	$\sigma_{R^q}(x)$	$b^q(x)$
R^1	(.33, .98)	.33	(.07, .04)
R^2	(0, .98)	0	(0, 0)
R^3	(.33, .02)	.02	(.01, 0)
R^4	(.67, .98)	.67	(.12, .07)
R^5	(.67, 0)	0	(0, 0)
R^6	(.33, 0)	0	(0, 0)
R^7	(0, 0)	0	(0, 0)
R^8	(.67, .02)	.02	(0, .01)
Grado de consistencia final			(**.12**, .07)
Asignación de clase			Clase C_1

Regla	Gr. Compat.	Gr. Emp.	Nuevo Gr. Asoc.
R^q	$\mu^q_{A_i}(x)$	$\sigma_{R^q}(x)$	$b^q(x)$
R^1	(.39, .51)	.39	(.07, .04)
R^2	(0, .51)	0	(0, 0)
R^3	(.39, 0)	0	(0, 0)
R^4	(.61, .51)	.51	(.09, .05)
R^5	(.61, .49)	.49	(.19, 0)
R^6	(.39, .49)	.39	(.12, 0)
R^7	(0, .49)	0	(0, 0)
R^8	(.61, 0)	0	(0, .01)
Grado de consistencia final			(**.19** , .05)
Asignación de clase			Clase C_1

Tabla 6.5: Proceso completo de inferencia. Los nuevos grados de asociación están referidos a aquellos obtenidos mediante el ajuste de los grados de las reglas. Instancia x_1 (arriba) y x_{251} (abajo)

tra en los grados de certeza de la BR, modificando la fuerza de asociación de las reglas con cada clase antes de la aplicación del proceso de inferencia. Por lo tanto, el enfoque local actúa después del proceso de aprendizaje de la base de reglas, mientras que el enfoque global solo actúa al final del MRD en los grados finales de pertenencia obtenidos de la muestra de entrenamiento.

En consecuencia, como se expone en [80], el modelo de post-proceso bipolar global es susceptible de ser aplicado directamente sobre cualquier algoritmo de clasificación *soft*, independiente de la forma en que se obtenga la información *soft* final (probabilidades, grados de consistencia difusos, etc.), y por lo tanto sus principales ventajas son su flexibilidad y amplia aplicabilidad. Por otro lado, el modelo de ajuste bipolar local esta diseñado específicamente para su aplicación en SCBRDs, y por lo tanto se encuentra restringido a este contexto. Sin embargo, como se muestra en este ejemplo y en los resultados experimentales de las Sección 6.2.3, el enfoque local pa-

rece permitir una mayor capacidad de corrección, lo que lleva a producir clasificadores con mayor rendimiento que el global.

6.2.3. Resultados

El objetivo de esta sección es presentar los resultados del estudio experimental descrito anteriormente, realizado para evaluar la capacidad de mejora del ajuste bipolar local propuesto con respecto al clasificador difuso de referencia. En primer lugar, comparamos los dos enfoques bipolares junto con el de método de referencia Chi, con el objetivo de evaluar el incremento en la capacidad de clasificación dado por la sinergia de los SCBRDs y las metodologías de representación del conocimiento bipolar.

La Tabla 6.7 contiene los resultados alcanzados por los tres algoritmos comparados en términos del promedio de kappa (media) junto con su desviación estándar (sd) a lo largo de las cinco particiones de prueba consideradas en este experimento. Queda claro al observar la tabla, que se produce un aumento importante en términos de la métrica kappa seleccionada, especialmente en el caso del enfoque bipolar local, que obtiene una mejora promedio de 0.062, con su máximo en 0.299 para el conjunto de datos *Iris*. En cuanto al post-proceso global, el incremento promedio de kappa es de 0.015, con su máximo en 0.198 para el conjunto de datos *Glass*.

La Figura 6.4 muestra la distribución de los resultados dados en términos de las métricas *kappa* y *Accuracy* para evaluar gráficamente el comportamiento de nuestras propuestas bipolares incluso en términos de tan conocida medida. Se observa el aumento de las dos métricas mencionadas desde el método *Referencia* hasta el enfoque más preciso de *bipAdd_ Local*.

Según este experimento, podemos extraer los siguientes comentarios.:

- El uso de una representación de conocimiento bipolar como un complemento en la etapa de toma de decisiones del proceso de clasificación resulta adecuado. Parece claro que ambos enfoques superan los resultados logrados por el SCBRD de referencia considerando la métrica kappa, que equilibra el rendimiento de las dos clases objetivo. Por lo tanto, este esquema bipolar parece proporcionar una ventaja competitiva en el escenario de clasificación.

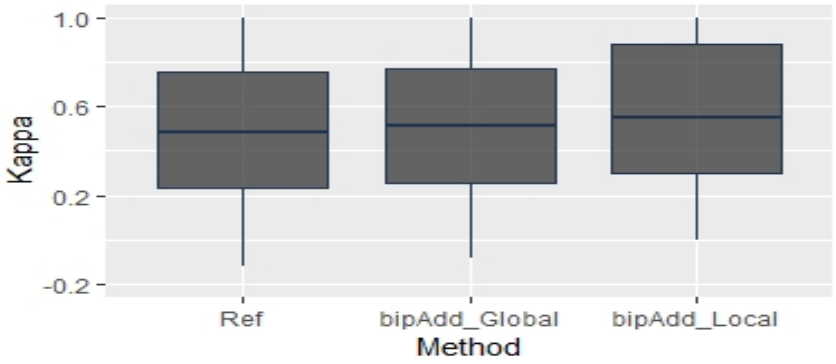
- El enfoque local propuesto parece ser mucho más robusto que el global. De hecho, los resultados obtenidos apoyan nuestra intuición con respecto a la necesidad de considerar un nivel más profundo de acción bipolar mediante el ajuste de los grados de disimilitud a un nivel de regla. Este esquema local alcanza una mejor adaptación a cada cluster y sus características inherentes.

- Los métodos bipolares propuestos superan al clasificador de referencia no solo en términos de la métrica kappa, sino también al considerar la

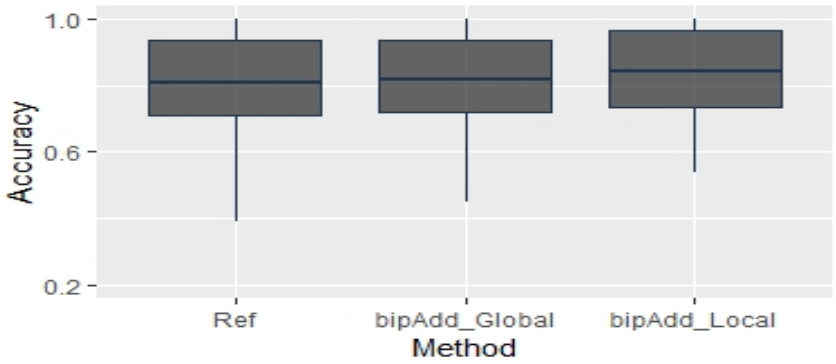
	CHI					
	Ref		bipAdd_Global		bipAdd_Local	
	Tr.	*Tst*	*Tr.*	*Tst*	*Tr.*	*Tst*
Aba	.096 ± .007	**.088 ± .046**	.099 ± .013	.076 ± .059	.096 ± .007	**.088 ± .046**
App	.487 ± .105	.366 ± .285	.661 ± .084	**.501 ± .343**	.692 ± .023	.474 ± .228
Aus	.821 ± .017	**.616 ± .057**	.828 ± .011	.610 ± .015	.847 ± .010	.606 ± .044
Bal	.544 ± .024	.538 ± .068	.544 ± .024	.538 ± .068	.590 ± .022	**.544 ± .113**
Ban	.107 ± .006	.104 ± .019	.117 ± .006	.111 ± .022	.262 ± .056	**.251 ± .086**
Bup	.139 ± .058	−.047 ± .076	.161 ± .037	.034 ± .026	.271 ± .049	**.116 ± .071**
Car	.870 ± .011	.721 ± .040	.901 ± .004	.785 ± .053	.901 ± .004	**.787 ± .052**
Con	.271 ± .027	.057 ± .031	.271 ± .027	.057 ± .031	.358 ± .026	**.164 ± .041**
Eco	.597 ± .015	.492 ± .106	.607 ± .006	.487 ± .093	.664 ± .028	**.545 ± .122**
Gla	.395 ± .090	.344 ± .199	.579 ± .050	**.542 ± .092**	.556 ± .046	.486 ± .101
Hay	1 ± .000	.929 ± .113	1 ± .000	.929 ± .113	1 ± .000	.929 ± .113
Hea	.957 ± .011	**.391 ± .048**	.963 ± .009	.345 ± .155	.963 ± .013	.337 ± .166
Iri	.707 ± .037	.701 ± .103	.766 ± .036	.641 ± .126	1 ± .000	**1 ± .000**
Led	.704 ± .092	.638 ± .174	.704 ± .092	.638 ± .174	.704 ± .092	.638 ± .174
Lin	.976 ± .009	.038 ± .063	.976 ± .009	**.151 ± .171**	.976 ± .009	**.151 ± .171**
Mag	.376 ± .033	.369 ± .016	.376 ± .033	.369 ± .016	.472 ± .040	**.465 ± .037**
Nty	.765 ± .034	.728 ± .169	.786 ± .034	.680 ± .194	.987 ± .012	**.946 ± .050**
Pag	.527 ± .053	.344 ± .267	.531 ± .052	.296 ± .259	.630 ± .066	**.479 ± .208**
Pen	.994 ± .004	.976 ± .029	.998 ± .003	.965 ± .039	.994 ± .004	**.976 ± .029**
Pho	.234 ± .009	.232 ± .050	.245 ± .009	.237 ± .053	.277 ± .013	**.260 ± .035**
Pim	.473 ± .018	.379 ± .096	.477 ± .015	.397 ± .070	.534 ± .017	**.415 ± .125**
Sah	.533 ± .058	.238 ± .110	.533 ± .058	.238 ± .110	.612 ± .020	**.276 ± .120**
Tit	.241 ± .118	.236 ± .101	.241 ± .118	.236 ± .101	.241 ± .118	.236 ± .101
Veh	.551 ± .026	.202 ± .070	.555 ± .024	**.226 ± .048**	651 ± .014	.222 ± .091
Vow	.948 ± .007	.917 ± .037	.968 ± .012	.903 ± .067	.981 ± .007	**.937 ± .040**
Win	.882 ± .024	.765 ± .072	.947 ± .017	.709 ± .074	.939 ± .035	**.806 ± .104**
Wis	.991 ± .004	**.791 ± .037**	.991 ± .004	**.791 ± .037**	.992 ± .004	.783 ± .011
Wnq	.359 ± .042	.266 ± .081	.387 ± .025	.302 ± .086	.480 ± .025	**.307 ± .073**
Wqw	.115 ± .018	.100 ± .012	.132 ± .011	.090 ± .018	.153 ± .005	**.135 ± .023**
Yea	.354 ± .015	.331 ± .053	.408 ± .022	.397 ± .093	.473 ± .018	**.448 ± .083**
Med.	.567 ± .301	.428 ± .303	.592 ± .300	.443 ± .293	.643 ± .289	**.494 ± .300**

Tabla 6.7: Resultados del experimento en los conjuntos de entrenamiento (*Tr.*) y prueba (*Tst.*) obtenidos por el clasificador Chi y las propuestas bipolares global y local. Métrica Kappa (Media ± Desviación típica)

métrica de precisión clásica para la comparación.



(a) kappa



(b) accuracy

Figura 6.4: Distribuciones de medidas de evaluación de los modelos bipolares frente a la referencia

Para detectar diferencias significativas entre los resultados de los diferentes clasificadores, llevamos a cabo el test de rangos alineados de Friedman, Figura 6.5. Es claro que existe una diferencia de rendimiento entre el enfoque bipolar local y los dos restantes clasificadores. La prueba de Friedman proporciona un p-valor cercano a cero, lo que implica que existen diferencias estadísticamente significativas entre los resultados proporcionados por cada método.

Por lo tanto, podemos comparar nuestra metodología novedosa con los enfoques restantes mediante la aplicación de una prueba Holm post-hoc utilizando el mejor enfoque (el que tiene una clasificación más baja) como método de control y el cálculo del p-valor ajustado (APV) para los dos métodos restantes.

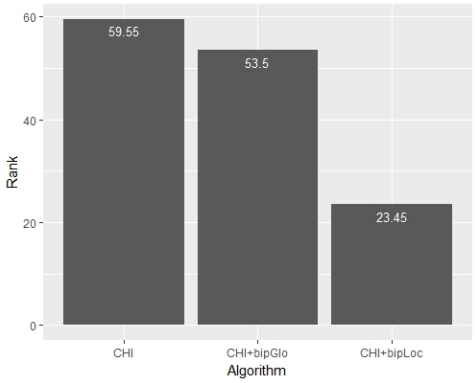


Figura 6.5: Rangos de los tres métodos comparados

La Tabla 6.8 contiene el resultado de la prueba post hoc de Holm, con un claro rechazo de la hipótesis nula de igualdad de rendimiento de los enfoques comparados tanto para el método de referencia de Chi como para el ajuste posterior bipolar global. Esto señala la superioridad del enfoque bipolar local cuando se compara no solo con la referencia, sino también con nuestra propuesta bipolar global anterior. Sin embargo, el post-ajuste global no parece proporcionar una mejora significativa con respecto a los resultados del método Chi de referencia.

i	Algoritmo	Hipótesis	APV
2	CHI	Rechazada for CHI+bipLoc	**0***
1	CHI+bipGlo	Rechazada for CHI+bipLoc	**.000008***

Tabla 6.8: Test de Holm para comparar el clasificador CHI+bipLoc frente al resto.

En resumen, a lo largo de este experimento una sinergia positiva entre SCBRDs y RBC se ha evidenciado, al menos para el clasificador de base estudiado (método de Chi). Además, esta aseveración encuentra un sólido apoyo estadístico.

Capítulo 7

Un nuevo sistema de explotación de información bipolar multidimensional basado en reglas

Una cosa es una cosa y otra cosa...son dos cosas.

Suko Peña

Resumen: Este capítulo se reserva enteramente al desarrollo y evaluación del novedoso sistema de explotación de pares de información bipolar basado en reglas brevemente introducido en la Sección 3.3.2. En primer lugar, se expone la propuesta y el marco teórico a ella asociado en la Sección 7.1. A continuación se presentan los principales resultados obtenidos en la Sección 7.2, finalizando el capítulo con un resumen de las lecciones aprendidas a modo de conclusiones en la Sección 7.2.

7.1. Explotación multidimensional basada en reglas

En esta sección se describe un nuevo esquema de explotación de las puntuaciones bipolares. Como es bien sabido, cualquier clasificador *soft* devuelve, tras su entrenamiento, un vector de puntuaciones $ev(x) = (ev_1(x), \dots ev_c(x))$ cuantificando la fuerza de asociación entre el patrón x y cada una de las clases. Por lo general, los clasificadores explotan estas puntuaciones aplicando la regla del máximo o seleccionando el umbral de curva ROC óptimo para producir la decisión o asignación de clase final.

En lugar de estas opciones, la propuesta aquí descrita, que puede ser considerada una particularización del marco de funciones de explotación detallado en la Sección 3.3.2, consiste en introducir primero la representación bipolar de las puntuaciones descrita en el Capítulo 3 para los casos probabilístico y difuso y posteriormente ajustar un clasificador basado en reglas como CART a estas puntuaciones bipolares como regla de decisión. Este mecanismo de decisión basado en la $RBC + CART$ se inserta en una búsqueda evolutiva de la matriz de disimilitud (ver Sección 2.8) para obtener la estructura de disimilitud que produce los mejores resultados.

Por lo tanto, el método propuesto funciona de la siguiente manera: primero, se ajusta un clasificador de base o de referencia a los datos de entrenamiento . Esto produce un vector $ev^+(x) = ev(x)$ de evidencia *soft* para cada patrón x en la muestra de entrenamiento . Luego, se lleva a cabo un proceso de optimización evolutivo de la matriz de disimilitud D. Esta búsqueda, a su vez, opera como se muestra en la Figura 7.1:

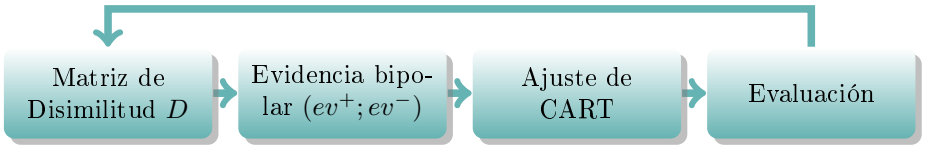


Figura 7.1: Proceso de optimización evolutivo de la matriz de disimilitud D basada en CART.

Para cada matriz D probada, las puntuaciones negativas $ev^-(x)$ se obtienen de las puntuaciones proporcionadas como se describe en la Sección 3.3.1, y luego se ajusta un árbol de clasificación CART (con parámetros predeterminados) usando los pares bipolares $(ev_i^+(x), ev_i^-(x))$, $i = 1, \ldots, c$, como características de entrada y las etiquetas de la clase de entrenamiento como respuesta. De esta manera, a cada instancia de entrenamiento se le asigna una clase predicha por el algoritmo CART, y se puede obtener una medida de rendimiento del entrenamiento para evaluar la matriz D que se está probando. En este trabajo, el rendimiento se mide utilizando el índice kappa detallado en la Sección 2.7.

El proceso de optimización evolutivo finalmente devuelve una matriz con mejor comportamiento, D^*, de entre la población de matrices inicial, así como una base de reglas ajustada por CART, que determina cómo deben clasificarse los patrones de acuerdo con los pares de evidencias bipolares, obtenidas tras aplicar el esquema de RBC con la matriz D^* a la evidencia *soft* proporcionada por clasificador de referencia. Esta base de reglas ajustada se puede aplicar para predecir nuevos patrones o instancias de prueba toda vez el clasificador base proporcione los grados difusos correspondientes, mejorando su poder predictivo y proporcionando una respuesta de mayor confianza al tener en cuenta la dicotomía bipolar.

Algoritmo 2 Procedimiento para la aplicación del esquema de explotación de información bipolar multidimensional basado en reglas

Seleccionar un clasificador soft base C_S, su configuración paramétrica p_{C_S} y el número de *folds* F.
for cada *fold* f $\in F$ **do**
  Ajustar el clasificador base al conjunto de entrenamiento y computar el vector de evidencia soft $ev(x)$ *para cada patrón* x*, con* $ev^+(x) = ev(x)$
  procedure PROCESO DE OPTIMIZACIÓN EVOLUTIVO DE LA MATRIZ DE DISIMILITUD D(EVO)
    for cada iteración del algoritmo evolutivo **do**
      Generar una matriz de disimilitud D.
      Calcular el vector de evidencia de carácter negativo $ev^-(x) = Dev^+(x)$ *para cada patrón* x.
      Construir el conjunto de información soft compuesto por la evidencia bipolar $(ev^+(x), ev^-(x))$ *y la clase real* $C(x)$ *para cada patrón* x.
      Ajustar el meta-clasificador al conjunto de información soft y obtener la clasificación ajustada final dada por la base de reglas.
      Computar el valor de kappa de la clasificación κ *en el conjunto de entrenamiento .*
    end for
    Seleccionar la matriz D^* *con mayor valor* κ *en el conjunto de entrenamiento .*
  end procedure
end for
Calcular el promedio de rendimiento en los conjuntos de entrenamiento y prueba a lo largo de los F *folds.*

Por lo tanto, el sistema toma la evidencia proporcionada por el clasificador base y, dada la matriz de disimilitud hallada en el proceso evolutivo, D^*, construye la evidencia negativa como se muestra en la Sección 3.2. Una vez que se determina los grados de evidencia negativa, en lugar de aplicar las funciones de agregación exploradas anteriormente (ver Sección 3.3.1), se utiliza un meta-clasificador basado en reglas como CART para decidir la clase final. Este método se puede enmarcar en el paradigma de explotación de información bipolar presentado en la Sección 3.3.2, donde ahora la función de explotación Φ viene dada por un meta-clasificador Θ.

Definición 7.1.1. Sean $ev^+ = (ev_1^+, ..., ev_c^+)$ y $ev^- = (ev_1^-, ..., ev_c^-)$ las evidencias positiva y negativa proporcionada por el clasificador base y Θ un meta-clasificador, entonces la clase determinada por el *Clasificador Bipolar* (en el sentido descrito en el Capítulo 3) para una instancia x, viene determinada por la predicción alcanzada por el meta-clasificador Θ aplicado sobre los vectores de evidencias positivas y negativas dados por el clasificador base para esa misma observación $ev^+(x) = (ev^+(x)_1, ..., ev_c^+(x))$ y

$ev^-(x) = (ev_1^-(x), ..., ev_c^-(x))$.

$$Clase(x) = C_i \text{ si y solo si } \Theta(ev^+(x), ev^-(x)) = C_i \tag{7.1}$$

La asignación de clase ajustada se determina mediante la aplicación del conjunto de reglas dadas por el meta-clasificador sobre las evidencias positivas y negativas.

La matriz de disimilitud se puede encontrar por medio de alguna de las metaheurísticas descritas en 2.8. En este trabajo se consideró un optimizador de lobos grises (GWO) [57].

7.2. Resultados experimentales

Esta sección está dedicada a proporcionar una comparación del rendimiento del método propuesto con el de los enfoques aditivos y logísticos introducidos en las Sección 3.3.1. Con este objetivo, se lleva a cabo un estudio experimental completo en el contexto de problemas de clasificación de tres clases. Por lo tanto, a partir de la evidencia *soft* dada por cada algoritmo base, y siguiendo la exposición presentada en las Secciones 3.3.1 y 3.3.2, se realizan tres búsquedas diferentes basadas en GWO, una para cada método evaluado (agregaciones aditiva y logística y explotación basada en CART), para obtener las mejores estructuras de disimilitud para cada método.

Con respecto a los clasificadores básicos, en este trabajo seleccionamos dos algoritmos de clasificación competitivos como *Random Forest* (RF) [9] y *eXtreme Gradient Boosting* (XGB) [13]. Los tres enfoques bipolares junto con el clasificador base se ajustan y se prueban en banco formado por 30 conjuntos de datos multiclase del repositorio de conjunto de datos KEEL [75], utilizando un marco experimental de validación cruzada 5-*folds*. Para transformar conjuntos de datos multiclase en conjuntos de tres clases, hemos tomado como clase C_1 y C_2 los originales más cercanos al 20 % de las instancias, y como clase C_3 la unión de las clases restantes. Finalmente, se aplican pruebas estadísticas para evaluar rigurosamente los resultados.

La Tabla 7.1 resume la información sobre los seleccionado conjuntos de datos, mostrando en primer lugar las categorías escogidas como C_1 y C_2, el número de ejemplos (#Ex.), el número de atributos (#Atts.) y su tipo (Real/Integer/Nominal).

El objetivo es, por lo tanto, comprobar si el modelo de representación bipolar basado en disimilitudes permite una mejora del rendimiento de los clasificadores básicos, bajo diferentes estrategias de explotación y para diferentes clasificadores básicos.

Las Tablas 7.2 y 7.3 muestran, para cada conjunto de datos, el índice kappa medio de cada método a lo largo de los cinco conjuntos de prueba. La

Id.	**Data-set**	**Clase C_1**	**Clase C_2**	**#Ex.**	**#Atts.**	**(R/I/N)**
Aba	Abalone	9	10	4174	8	(7/1/0)
Aut	Autos	0	1	690	14	(3/5/6)
Bal	Balance	L	R	625	4	(4/0/0)
Car	Car	unacc	acc	159	25	(15/0/10)
Cle	Cleveland	0	1	303	13	(13/0/0)
Con	Contraceptive	1	3	1473	9	(6/0/3)
Der	Dermatology	1	3	366	34	(0/34/0)
Eco	Ecoli	cp	im	336	7	(7/0/0)
Fla	Flare	H	C	1066	25	(15/0/10)
Gla	Glass	2	1	214	9	(9/0/0)
Hay	Hayes-Roth	1	2	160	4	(0/4/0)
Iri	Iris	Setosa	Versicolor	150	4	(4/0/0)
Led	Led7digit	3	7	500	7	(7/0/0)
Lin	Lymphography	metastases	malign_lymph	148	18	(3/0/15)
Nty	Newthyroid	1	2	215	5	(4/1/0)
Nur	Nusery	not_recom	priotity	12690	8	(0/0/8)
Pag	Page-blocks	1	2	5472	10	(4/6/0)
Pen	Penbased	0	1	10992	16	(0/16/0)
Sat	Satimage	1	7	462	9	(5/3/1)
Seg	Segment	1	2	2310	19	(19/0/0)
Shu	Shuttle	1	4	2175	9	(9/0/0)
Spl	Splice	N	IE	3190	60	(0/0/60)
Tae	Tae	3	2	151	5	(0/5/0)
Thy	Thyroid	3	2	720	21	(6/0/15)
Veh	Vehicle	bus	van	846	18	(0/18/0)
Vow	Vowel	0	10	990	13	(10/3/0)
Win	Wine	2	1	178	13	(13/0/0)
Wqr	Winequality-red	5	6	1599	11	(11/0/0)
Yea	Yeast	CYT	MIT	1484	8	(8/0/0)
Zoo	Zoo	1	2	101	16	(0/0/16)

Tabla 7.1: Descripción de los conjuntos de datos empleados en la propuesta de explotación multidimensional basada en reglas.

explotación basada en CART obtiene los mejores resultados en términos de promedio global de kappa en los conjuntos de prueba para RF y XGB.

Dado que los clasificadores básicos logran un rendimiento perfecto en muchos conjuntos de entrenamiento , las agregaciones tanto aditiva como logística no encuentran margen para mejorar su rendimiento, y por tanto alcanzan una estructura de disimilitud nula ($D = 0$). Sin embargo, la fortaleza de esta novedosa explotación basada en CART reside en su capacidad para superar los resultados básicos en prueba incluso en este difícil escenario. De hecho, incluso considerando la matriz de disimilitud nula, se ajusta un árbol CART considerando solo las reglas con respecto a la evidencia positiva. Esta base de reglas se aplicará para tomar la decisión final en el conjunto de prueba, lo que permite mejorar los resultados de referencia.

	RF							
	Ref		bipAdd		bipLog		bipCart	
	Tr.	*Tst.*	*Tr.*	*Tst.*	*Tr.*	*Tst.*	*Tr.*	*Tst.*
Aba	1	.112	1	.112	1	.112	1	**.170**
Aut	1	.698	1	.698	1	.698	1	.698
Bal	.616	.540	.616	.540	.616	.540	.616	.540
Car	.995	.869	1	.869	.995	.869	.995	.869
Cle	.730	.501	.948	**.506**	.730	.501	.980	.501
Con	.788	.269	.814	**.284**	.800	.278	.821	.282
Der	1	.995	1	.995	1	.995	1	.990
Eco	1	.767	1	.767	1	.767	1	.767
Fla	.797	.785	.807	.781	.797	.785	.797	.785
Gla	1	.673	1	.673	1	.673	1	.673
Hay	.886	.715	.886	**.721**	.886	.718	.886	**.721**
Iri	1	.953	1	.953	1	.953	1	**.963**
Led	.752	.678	.775	.687	.776	.687	.775	**.703**
Lin	.990	.681	1	.683	.996	.710	.990	**.724**
Nty	1	.931	1	.931	1	.931	1	**.938**
Nur	.999	.983	1	.983	.999	.983	.999	.983
Pag	1	.828	1	.828	1	.828	1	**.833**
Pen	1	.906	1	.906	1	.906	1	.906
Sat	1	.907	1	.907	1	.907	1	.910
Seg	1	.993	1	.993	1	.993	1	.995
Shu	1	.996	1	.996	1	.996	1	.996
Spl	1	.598	1	.598	1	.598	1	**.698**
Tae	.953	.467	.954	.467	.953	.467	.953	.467
Thy	1	.909	1	.909	1	.909	1	.909
Veh	1	.950	1	.950	1	.950	1	.950
Vow	1	.862	1	.862	1	.862	1	.862
Win	1	.972	1	.972	1	.972	1	.972
Wnq	1	.528	1	.528	1	.528	1	.528
Yea	1	.362	1	.362	1	.362	1	**.385**
Zoo	1	1	1	1	1	1	1	1
Mean	.950	.748	.960	.749	.952	.749	.960	**.757**

Tabla 7.2: Resultados obtenidos por el clasificador Random Forest (RF). Promedio de la métrica kappa.

	XGB							
	Ref		bipAdd		bipLog		bipCart	
	Tr.	*Tst.*	*Tr.*	*Tst.*	*Tr.*	*Tst.*	*Tr.*	*Tst.*
Aba	.919	.134	.958	.152	.925	.177	.919	**.179**
Aut	1	.734	1	.734	1	.734	1	**.762**
Bal	.605	.569	.612	.577	.605	**.588**	.610	.573
Car	1	.986	1	.986	1	.986	1	.986
Cle	.642	.546	.819	.496	.669	.392	.642	.546
Con	.415	.308	.456	**.311**	.415	.308	.415	.308
Der	1	.993	1	.993	1	.993	1	.982
Eco	.925	.744	.940	.704	.925	.744	.925	.744
Fla	.795	.789	.802	.787	.799	.787	.797	**.791**
Gla	.997	.652	1	.617	1	.652	.987	**.673**
Hay	.874	.708	.882	.683	.874	**.743**	.874	.717
Iri	1	.937	1	.937	1	.937	1	**.943**
Led	.755	.667	.767	.675	.755	.676	.770	**.689**
Lin	.961	.643	.970	.686	.961	**.698**	.961	.688
Nty	.999	.913	1	.916	1	**.927**	1	.913
Nur	1	.998	1	.997	1	.997	1	.998
Pag	.974	.805	.992	**.835**	.975	.812	.992	.807
Pen	1	.915	1	.915	1	.915	1	**.916**
Sat	1	.919	1	.919	1	.919	1	.918
Seg	1	.994	1	.994	1	.994	1	.995
Shu	1	.990	1	.990	1	.990	1	.990
Spl	1	.552	1	.552	1	.552	1	**.553**
Tae	.938	.373	.944	.355	.938	.373	.944	**.417**
Thy	1	.866	1	.866	1	.866	1	**.877**
Veh	1	.954	1	.954	1	.954	1	.954
Vow	1	.903	1	.903	1	.903	1	.903
Win	1	.961	1	.961	1	.961	1	**.980**
Wnq	.999	.504	1	**.511**	.999	.506	.999	.506
Yea	.632	.364	.663	**.389**	.632	.365	.648	.387
Zoo	1	1	1	1	1	1	1	1
Mean	.914	.747	.927	.747	.916	.748	.916	**.757**

Tabla 7.3: Resultados obtenidos por el clasificador eXtreme Gradient Boosting (XGB). Promedio de la métrica kappa.

Algoritmo	Rango RF	Rango XGB
Ref	71.85	72.73
BipAdd	64.90	66.48
BipLog	66.38	61.16
BipCart	**38.87**	**41.62**
Friedman p-val	.000034	.000029
Holm APV (Mejor vs. Ref)	**.00072**	**.00159**

Tabla 7.4: Rangos promedio (Aligned Friedman), p-valores asociados y APV del test de Holm para cada algoritmo.

Comparación	R^+	R^-	p-val
RFbipCart vs. RFRef	358.0	107.0	**.0094**
RFbipCart vs. RFbipAdd	305.5	129.5	**.0523**
RFbipCart vs. RFbipLog	356.0	109.0	**.0104**
XGBbipCart vs. XGBRef	403.5	61.5	**.00024**
XGBbipCart vs. XGBbipAdd	315.0	120.0	**.0238**
XGBbipCart vs. XGBbipLog	310.5	154.5	.0878

Tabla 7.5: Test de Wilcoxon para comparar la nueva propuesta basada en $RBC + CART$ (R^+) frente a las restantes aproximaciones (R^-).

Con el fin de proporcionar soporte estadístico a estos hallazgos, se realizó un test de rangos alineados de Friedman para comparaciones múltiples, considerando los tres enfoques junto con la referencia. Esta prueba muestra p-valores cercanos a cero para los clasificadores RF y XGB, lo que significa que hay diferencias significativas entre los resultados. Además, se aplica la prueba post-hoc de Holm para contrastar qué alternativas son superadas por el método de control (es decir, nuestra propuesta de explotación basada en CART) y la referencia en un marco de comparación múltiple. La Tabla 7.4 refleja las clasificaciones promedio para cada enfoque junto con el p-valor ajustado de Holm (APV) para cada clasificador de referencia. Considerando el APV bajo para ambos algoritmos base, se puede concluir que el método propuesto logra una mejora significativa.

Además, para comparar el comportamiento del método propuesto con los otros en un esquema de pares, se realiza una prueba de Wilcoxon. Nuevamente, los p-valores en la Tabla 7.5 apoyan el rechazo de la hipótesis nula de igualdad en los resultados para cada par de métodos considerados con un nivel de confianza de 90 %. Por lo tanto, el nuevo enfoque supera no solo a la referencia, sino también a los otros dos clasificadores bipolares.

7.3. Lecciones aprendidas

En esta sección, se propuso un nuevo esquema de explotación basado en CART para aprovechar los efectos positivos y negativos dados por el paradigma de representación del conocimiento bipolar en el marco de la clasi-

ficación multi-clase. Se han realizado varias comparaciones para evaluar el comportamiento de este nuevo método que evidencia una mejora significativa en los resultados de este enfoque con respecto no solo a los clasificadores *soft* de referencia sino también al resto de dos métodos bipolares, aditivo y logístico.

Teniendo en cuenta los resultados obtenidos en el estudio experimental, este nuevo modelo de explotación parece ser una solución adecuada para manejar clasificadores de alta precisión como *Random Forest* (RF) y *eXtreme Gradient Boosting* (XGB) [13] debido a su aptitud para mejorar los resultados en la muestra de prueba cuando la referencia alcanza un ajuste perfecto en el conjunto de entrenamiento .

Cabe destacar que cualquier meta-clasificador podría ser considerado en este punto, si se pretende fomentar la interpretabilidad de la función de explotación, la elección de cualquier tipo Clasificador Basado en Reglas (CBR) resulta adecuada. Se detallan en la Sección 8.3 las posibles extensiones en esta línea de investigación.

Capítulo 8

Conclusiones y Trabajo futuro

Podemos ver poco sobre el futuro, pero lo suficiente para darnos cuenta de que hay mucho que hacer.

Alan Turing

Resumen: En este capítulo se detallan las principales conclusiones del trabajo presentado en esta memoria bajo distintas perspectivas. En la Sección 8.1 se destacan, por un lado, los aspectos más relevantes en cuanto a la consideración de un marco de toma de decisiones que trascienda la extendida regla del máximo así como una estructura de relaciones en el conjunto de clases. Se resalta, por otro lado, la importancia de obtener reglas de clasificación dadas en términos del lenguaje natural y como, la utilización de Sistemas Difusos Evolutivos (EFS) desarrollados como extensión de los SCBRDs, cobran cada día mayor importancia por su alta explicabilidad en el marco de la Inteligencia Artificial Explicable. En este sentido se pone de manifiesto la sinergia positiva producida entre este tipo de clasificadores y la RBC. Desde el punto de vista de cumplimiento de los objetivos propuestos, en la Sección 8.2, se lleva a cabo un esquema de relaciones entre las propuestas realizadas en cada objetivo marcado destacando las contribuciones que se derivan de ellas. Para finalizar, se proponen en la Sección 8.3 varias de las posibles líneas futuras de investigación a las que puede dar pie este trabajo.

8.1. Principales conclusiones

En este apartado se pretende hacer visibles las principales enseñanzas extraídas del proceso de elaboración de este trabajo en un sentido filosófico.

Una vez se ha entendido la motivación subyacente a todo este trabajo, que no es más que la de una creencia en la posibilidad de explotar la información intermedia de los algoritmos mediante aproximaciones más flexibles que la conocida regla del máximo, es inmediato deducir el sentido en que aquí se entienden los procesos relativos a la toma de decisiones en el marco de la clasificación en MDD.

En contraposición al creciente fervor provocado por los algoritmos de *machine learning* de tipo "caja negra" como las Redes Neuronales Convolucionales, en general, los modelos del ámbito conocido como *Deep Learning* [93], se aboga aquí por la utilización de sistemas para la toma de decisiones que se caractericen por su interpretabilidad. Es más, en cierta parte, se pretende que esta interpretabilidad sea entendida en términos habitualmente manejados por el ser humano.

En ningún caso esta determinación supone desprecio alguno por los modernos algoritmos de clasificación supervisada que tan alto rendimiento han demostrado en el contexto de la MDD, de hecho, el método propuesto se construye enteramente bajo la premisa de una clasificación de base realizada por alguno de ellos. A lo largo de este trabajo se hace uso de algunos de estos algoritmos como Random Forest o Redes Neuronales y, en todo momento, se pretende generar sinergias entre la representación bipolar del conocimiento y los clasificadores considerados por medio de la aplicación de la primera en dos distintos niveles (global y local).

Conviene recordar que el método local propuesto está concebido para su aplicación en el contexto de SCBRDs, mientras que el enfoque global se erige como una alternativa general y flexible, susceptible de ser aplicada sobre cualquier algoritmo de naturaleza *soft*. Sin embargo, la consideración del paradigma local permite una representación más fiel de la información contenida en el propio entrenamiento del clasificador base, con lo que la RBC tiene mayor capacidad de ajustar las evidencias en un nivel de reglas. Como consecuencia, las mejoras más acusadas se obtienen mediante la aplicación de este paradigma bipolar local.

A lo largo de este tiempo de trabajo en las líneas de investigación detalladas en esta memoria, gran cantidad de aprendizajes han sido interiorizados mas muchos otros quedaron en el tintero como se comentará en próximas secciones. Se destacan aquí aquellas conclusiones que han sido fehacientemente probadas y demostradas en este trabajo.

En primer lugar, el enfoque global con el método de agregación de evidencias positiva y negativa ha sido aplicado a clasificadores probabilísticos enfrentando problemas tanto binarios como multiclase con algunas mejoras significativas en términos estadísticos. Este paradigma se entiende como un post-proceso y sus resultados finales son absolutamente dependientes de aquellos obtenidos por el algoritmo base considerado. Así, por ejemplo, un clasificador base con máximo rendimiento (digamos kappa = 1) en el con-

junto de entrenamiento pero no tan bueno en el conjunto de prueba, no es susceptible de ser ajustado por la RBC para la mejora de su capacidad de generalización a nuevas instancias, al menos, bajo el paradigma de aprendizaje + agregación aquí considerado. Por otra parte, el esquema de entrenamiento paramétrico del clasificador base considerado, lo hace dependiente de ciertos factores incontrolados obteniendo resultados, en ocasiones, faltos de la necesaria robustez.

Debido a estos escollos encontrados en el camino, se exploran sendos métodos para soslayar cada uno de las dos limitaciones comentadas: limitación en aprendizaje y falta de robustez. En el primer caso, se propone un sistema de explotación multidimensional de pares de evidencias bipolares basado en reglas como alternativa a los métodos agregativos estudiados, con resultados que evidencian su mayor rendimiento y capacidad de adaptación incluso en los casos de limitación extremos, lo que alienta un estudio riguroso en este sentido. En el segundo caso, de cara a minimizar el efecto de dependencia de factores incontrolados sobre el aprendizaje de los clasificadores probabilísticos, se propone en este trabajo un paradigma de agregación difusa de clasificadores probabilísticos que mejora su estabilidad, como antesala para la aplicación de la RBC en entorno difuso. En este punto se exploran tres distintos operadores de agregación (máximo, mínimo y media aritmética), con mejoras significativas en gran parte de las pruebas.

Cuando se explora el campo de los clasificadores de tipo difuso, la aplicación de la RBC se realiza en dos niveles distintos. En primer lugar, y como extensión de los métodos aplicado anteriormente en contexto probabilístico, se lleva a cabo el post-proceso bipolar a nivel global en base a la información dada por el algoritmo base en la etapa previa a la asignación final a la clase. En este caso, se consideran clasificadores difusos robustos obtenidos previamente así como SCBRD. Se confirman las bondades del post-proceso bipolar global también para este tipo de clasificadores soft.

Enfocando todos los esfuerzos en los SCBRDs y, por tanto, restringiéndonos a ellos, se desarrolla el enfoque local o a nivel de reglas comentado que puede entenderse como un ajuste bipolar de los pesos de las reglas en la etapa de aprendizaje del propio SCBRD. En este punto, se demuestra que la aplicación de la RBC a este nivel local para la creación de pesos de reglas bipolares obtiene la mayor tasa de mejora de entre todas las propuestas aquí presentadas.

Conclusión *filosófica* final 1: Sea cual sea la perspectiva desde la que se entiendan y aborden los problemas de clasificación, es claro que, la confianza inspirada por la decisión nítida realizada por una máquina no debería ser muy elevada, máxime cuando se desconoce el mecanismo generador de tal decisión. Se evidencia la necesidad de tomar decisiones en base a la información de la que dispone el algoritmo de un modo más amplio y flexible trascendiendo, de esta forma, la conocida regla del máximo.

Conclusión *filosófica* final 2: La consideración de una estructura de relaciones entre los grupos o clases se demuestra capaz de representar con mayor fidelidad la realidad subyacente a un problema de clasificación.

Conclusión *filosófica* final 3: Se evidencia la existencia de una fuerte sinergia entre los dos modelos de representación del conocimiento inspirados en el ser humano, los SCBRDs y la RBC.

8.2. Vínculos entre propuestas, objetivos y contribuciones

En este punto, se entiende necesaria la elaboración de un esquema de vínculos que aclare las relaciones existentes entre las propuestas realizadas en este trabajo y los objetivos inicialmente planteados, de cara a la evaluación del cumplimiento de los últimos. Más allá, resulta de igual forma relevante la vinculación de estos últimos pares con las contribuciones que se extraen de esta memoria, ya que componen de alguna manera, el soporte de este trabajo por parte de la comunidad investigadora.

En este sentido, se pretende representar de forma concisa este conjunto de relaciones siguiendo, en pos de la claridad, la estructura de orden presentada en la Sección 1.2 de objetivos de esta memoria. Por tanto, se distingue entre las aportaciones realizadas en relación a cada uno de los objetivos definidos.

Objetivo 1: *Explotar la información de naturaleza soft dada por cualquier algoritmo de clasificación supervisada en la etapa previa a la toma de decisiones, definiendo un nuevo marco para la toma de decisiones basado en la Representación Bipolar del Conocimiento (RBC).*

En el marco de este objetivo, has sido muchos los esfuerzos realizados para la proposición de distintos paradigmas de explotación de la información de naturaleza soft, dados por cualquier algoritmo de clasificación que pueda ser enmarcado en el conjunto de clasificadores de esta naturaleza. Concretamente, según los objetivos específicos planteados, se tiene:

1. *Definir formalmente el concepto de estructura de disimilitud entre grupos de instancias pertenecientes a distintas clases.*

   Como primer paso, se alienta la consideración de estructuras de disimilitud basadas en matrices. De esta forma, siguiendo la propuesta incluida en [73], se modelan las estructuras de relación entre clases por medio de estas conocidas estructuras matemáticas. Este esquema permite una representación de los conceptos relacionados con el antagonismo y la antonimia semánticos en el contexto de la toma de

decisiones. Esta propuesta se encuentra incluida en la Sección 3.1 del Capítulo 3.

2. *Estudiar las relaciones existentes entre estructuras de disimilitud bajo distintos tipos de agregaciones.*

   Tras la consideración de la estructura de disimilitud comentada, se hace necesario un profundo estudio sobre su comportamiento bajo distintos paradigmas bipolares. Así, en pos de un mejor entendimiento de tales estructuras, se realiza un análisis que arroja ciertas bondades, restricciones y limitaciones de las consideradas matrices en el contexto de distintas agregaciones. Este estudio permite extraer interesantes conclusiones sobre la capacidad de ajuste de los modelo bipolares propuestos, en concreto, en un contexto probabilístico. En un siguiente paso, se ha pretendido proponer un sistema de construcción de matrices de disimilitud como extensión del conocido ajuste mediante curvas ROC tan ampliamente aplicado en clasificación binaria. Así, el objetivo aquí es la generación automática de estructuras de disimilitud bajo este marco, extendiendo la metodología ROC a clasificación multiclase bajo alguno de los paradigmas de descomposición binaria: *One vs. One* o *One vs. All*. Los pormenores de estos análisis se encuentran incluidos en las Secciones 3.4 y 3.4.3 del Capítulo 3, respectivamente.

3. *Plantear un sistema de toma de decisiones global "a posteriori"basado en información bipolar aplicable a cualquier tipo de algoritmo de clasificación probabilístico o difuso.*

   Tras la realización de las propuesta y estudios anteriormente señalados, estamos en disposición de realizar las primeras propuestas formales, sometiéndolas al juicio de la comunidad investigadora. Concretamente se pueden dividir las contribuciones realizadas en dos grandes grupos:

   a) Marco Probabilístico: Las principales contribuciones en este contexto se detallan a continuación:

   Contribución 2: Villarino, G., Gómez, D., Rodríguez, J. T. (2017) Improving Supervised Classification Algorithms by a Bipolar Knowledge Representation. Advances in Intelligent Systems and computing 643 518–529. doi:10.1007/978-3-319-66827-7_48.
   Los resultados de esta aportación se muestran en la Sección 4.2.2 del Capítulo 4.

   Contribución 5: Villarino, G., Gómez, D., Rodríguez, J.T. (2018) Assessing the performance of bipolar classifiers in three-class problems. In Proceedings of CAEPIA Congress 2018 306–314.

Los resultados de esta aportación se muestran en la Sección 4.2.3 del Capítulo 4.

Contribución 7: Villarino, G., Gómez, D., Rodríguez, J.T., Fernández, A. (2019) A new exploitation scheme in the context of bipolar classifiers. ESCIM 2019 Congress. In press. Los resultados de esta aportación se muestran en Capítulo 7.

b) Marco Difuso: En lo referente a clasificadores difusos, existe cierto solapamiento entre este objetivo específico y los consiguientemente presentados, eligiendo enmarcar las contribuciones realizadas en éstos últimos.

4. Proponer un paradigma de clasificación bipolar robusta basado en replicación y agregación difusa en el contexto de la clasificación probabilística.

   En lo relativo a este objetivo específico, cabe destacar que la intención de éste es, de alguna forma, doble. Por un lado, se propone un nuevo sistema de generación de clasificadores difusos robustos a partir de aquellos probabilísticos. En otro plano, se pretende evaluar la capacidad de ajuste del paradigma bipolar a nivel global recientemente discutido, esta vez, en el contexto de los clasificadores de tipo difuso.

   Contibución 4: Villarino, G., Gómez, D., Rodriguez, J.T. et al. (2018) A bipolar knowledge representation model to improve supervised fuzzy classification algorithms. Soft Compututing **22** 5121–5144. doi:10.1007/s00500-018-3320-9.

   Los resultados de esta aportación se muestran en el Capítulo 5.

Objetivo 2: *Proponer una extensión del marco general de la inferencia en Sistemas de Clasificación Basados en Reglas Difusas (SCBRDs) basada en Representación Bipolar del Conocimiento (RBC).*

En el contexto de los SCBRDs, se aporta en este trabajo un esquema generalizado del método Inferencia Difusa enmarcado en el MRD. El principal objetivo de tal extensión es dotar al SCBRD de una capacidad añadida para manejar estructuras de RBC dadas por pares de evidencias positiva y negativa. En este caso no solo se considera el esquema global evaluado en anteriores propuestas, sino también enfoque local de la RBC. Se entiende el término nivel local en este caso como la etapa de entrenamiento del clasificador, concretamente el nivel de creación de la Base de Reglas.

1. *Extender el sistema de ayuda a la decisión global al contexto de los sistemas de clasificación basados en reglas (SCBRDs).*

Como primer método bipolar en el marco de los SCBRDs, se particulariza el marco general de RBC a nivel global descrito en el Capítulo 3. En este caso (ver Sección 6.1), la información de entrada para el post-proceso bipolar está compuesta por los grados de consistencia de la clasificación final.

2. *Proponer un nuevo mecanismo de representación bipolar del conocimiento a nivel local o nivel de reglas en el marco de los SCBRDs.*

   En este apartado se propone un esquema bipolar a nivel de reglas, entendido como un sistema de ajuste de los grados de certeza o pesos de las reglas bajo el paradigma de la RBC. Los pormenores a este respecto se encuentran en la Sección 6.1 del Capítulo 6.

3. *Extender el Método de Razonamiento Difuso (MRD) de los SCBRDs para el manejo de información de carácter bipolar.*

   No solo se aplica en este trabajo el esquema de clasificación de tipo nítido. Por el contrario, se propone una extensión bipolar del MRD que deja la puerta abierta a otros tipos de toma de decisiones o utilización de otras lógicas para la explotación de la información generada. Esta extensión se encuentra recogida en la Sección 6.1.4 del Capítulo 6.

   Los contenidos de este objetivo se recogen por completo en la siguiente contribución:

   Contribución 8: Villarino, G., Rodríguez, J.T., Gómez, D., Fernández, A. (2019): Extending the Fuzzy Inference by Bipolar Representation in a Classification Context: New Methods and Models.. [En preparación]

Objetivo 3: *Casos prácticos: por una parte, se pretende aplicar algunos de los avances al estudio de datos de siniestralidad vial proporcionados por la Dirección General de Tráfico (DGT), por otra, se considera la representación de la información soft en el caso especial de clasificación de segmentos en el contexto de la detección de bordes en imágenes.*

Contribución 1: Villarino, G., Gómez, D., Cintas, R., Rodriguez, J.T. (2016) Metodología de minería de datos para el estudio de tablas de siniestralidad vial. In Proceedings of CAEPIA-STYLF Congress 2016 599–608.

Contribución 3: Flores-Vidal, P. A., Gómez, D., Montero, J., Villarino, G. (2017) Classifying segments in edge detection problems. 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Nanjing, 2017, 1–6. doi:10.1109/ISKE.2017.8258764.

Contribution 6: Flores-Vidal, P. A., Villarino, G., Gómez, D., Montero, J. (2019) Classifying segments in edge detection problems. International

Journal of Computational Intelligence Systems **12** (1) 367–378.
doi:10.2991/ijcis.2019.125905653.

8.3. Una ventana al futuro

Todo camino tiene su final y es por ello que no han podido ser abordadas en este tiempo de investigación cuantas cuestiones se hubiera deseado. En efecto, muchas son las que quedan en el tintero, esperando pacientemente ser consideradas para su desarrollo formal.

De una parte, en lo que concierne a los sistemas de explotación de información de carácter bipolar, un amplio espectro de posibilidades son susceptibles de aplicación. Con respecto a la investigación futura sobre este enfoque, una línea principal de trabajo está constituida por el estudio de mecanismos adicionales que trasciendan las agregaciones aditivas y logísticas para explotar los pares de evidencia bipolar positiva y negativa. Una posibilidad particularmente atractiva es usar estos pares bipolares como la información base de una lógica multi-valuada como se explora en [62, 67, 77]. Esto permitiría un marco representativo aún más expresivo para aprovechar toda la información contenida en las puntuaciones *soft* proporcionadas por los clasificadores.

En este marco se sugiere, así mismo, el desarrollo de otra clase de sistemas de explotación no basados en agregación de los pares bipolares enmarcados en la filosofía presentada en la Sección 3.3.2. En este sentido, y en base a los buenos resultados obtenidos, parece interesante la consideración de otros sistemas de explotación basados en reglas difusas como los SCBRDs tipo FARC-HD [3], IVTURS [73] o FURIA [45], debido a su capacidad para generar una base de reglas interpretables a nivel lingüístico que determinen la clasificación final.

De cara a la consideración de un paradigma local en otros clasificadores de tipo probabilístico, se plantea la posibilidad de adición de un marco de RBC en la construcción del árbol de decisión, creando de esta forma un algoritmo de árbol bipolar que tenga en cuenta esta filosofía como criterio a la hora de realizar la particiones recurrentes propias de los arboles de decisión. Con ello se pretende enriquecer el conocimiento del algoritmo de clasificación introduciendo la información de carácter negativo, es decir, la probabilidad de pertenencia a las clases disimiles. Otra de las líneas interesantes es la generación de un método de RBC a nivel local para el algoritmo RF construido a partir de la aplicación del paradigma global a cada árbol individual.

El trabajo futuro relacionado con la sinergia entre SCBRD y la bipolaridad puede estar centrado en el análisis del comportamiento del enfoque local propuesto en clasificadores difusos básicos más precisos, como FARC-HD o FURIA. Además, también se explorarán diferentes enfoques para enriquecer estos clasificadores a través de un marco de representación bipolar, con

especial atención a cómo introducir la bipolaridad en las etapas de aprendizaje de estos modelos. Especial interés presenta la evaluación de distintos métodos para explotar la información bipolar en el marco del MRD ampliado aquí propuesto, así como el desarrollo de mecanismos más eficientes de aprendizaje de la estructura de disimilitud en el conjunto de clases. Una de las líneas de mayor relevancia en este campo, es el desarrollo de un marco de bipolaridad a nivel local-profundo (*Deep Local Approach*), en el que la estructura de disimilitud se aprenda de forma individual para cada regla, es decir, un aprendizaje centrado en los subespacios de cobertura de cada regla.

En un plano teórico, resulta de gran interés el estudio formal de las estructuras de disimilitud y su relación con la capacidad del mecanismo bipolar de producir cambios en la clasificación. De esta forma, se hace necesaria una rigurosa extensión del estudio parcial presentado en la Sección 3.4, esta vez, considerando todos los posibles marcos y realidades de aplicación.

Bibliografía

Y así, del mucho leer y del poco dormir,
se le secó el celebro de manera que vino
a perder el juicio.

Miguel de Cervantes Saavedra

[1] Abe, S. and Lan, M.S. (1995) A method for fuzzy rules extraction directly from numerical data and its application to pattern classification, *IEEE Transactions on Fuzzy Systems*, **3** 18–28.

[2] Adadi, A., Berrada, M. (2018) Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6 52138–52160

[3] Alcalá-Fdez, J., Alcalá, R., Herrera, F. (2011) A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning, *IEEE Transactions on Fuzzy Systems*, **19** (5) 857–872.

[4] Alcalá , R., Alcalá-Fdez J., Herrera, F. (2007) A Proposal for the Genetic Lateral Tuning of Linguistic Fuzzy Systems and Its Interaction With Rule Selection. *IEEE Transactions on Fuzzy Systems*, **15** (4) 616–635.

[5] Amo, A., Montero, J., Molina, E. (2001) Representation of consistent recursive rules. *European Journal of Operational Research*, **130** 29–53.

[6] Atanassov, K. T. (1999) Intuitionistic fuzzy sets theory and applications, Physica-Verlag, Heidelberg; New York.

[7] Biswas, S., Bordoloi, M., Singh, H., Purkayastha, B. (2016) A Neuro-Fuzzy Rule-Based Classifier Using Important Features and Top Linguistic Features. *International Journal of Intelligent Information Technologies*, **12** 38–50.

[8] Breiman L. (1984) Classification and Regression Trees New York, NY: Kluwer Academic Publishers-

[9] Breiman L. (2001) Random Forests. *Machine Learning* **40** 5–32.

[10] Bustince, H., Fernandez, J., Mesiar, R., Montero, J., Orduna, R. (2010) Overlap functions. *Nonlinear Anal., Theory, Methods Appl.*, 72 (3?4) 1488–1499.

[11] Cacioppo, J., Gardner, W., Berntson, G. (1997) Beyond bipolar conceptualizations and measures: the case of attitudes and evaluative space. *Personality and Social Psychology Review* **1** 3–25.

[12] Casillas J., Cordón O., Herrera F., Magdalena L. (2003) Interpretability Improvements to Find the Balance Interpretability-Accuracy in Fuzzy Modeling: An Overview. In: Casillas J., Cordón O., Herrera F., Magdalena L. (eds) Interpretability Issues in Fuzzy Modeling. Studies in Fuzziness and Soft Computing, (128).

[13] Chen, T., Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. KDD 2016.

[14] Chi, Z., Wu, J., & Yan, H. (1995) Handwritten numerical recognition using self-organizing maps and fuzzy rules. *Pattern Recognition* **28** (1) 59–66.

[15] Cohen J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20** 37–46.

[16] Cordón, O. Del Jesús, M.J., Herrera, F. (1999) A proposal on reasoning methods in fuzzy rule-based classification systems, *International Journal of Approximate Reasoning*, **20** (1) 21–45.

[17] Cutello, V., Montero, J. (1999) Recursive connective rules, *International Journal of Intelligent Systems*, **14** (1) 3–20.

[18] Demsar J., (2006) Statistical comparisons of classifiers over multiple datasets, *Journal of Machine Learning Research*, **7** 1–3.

[19] Dubois, D., Prade, H. (2002) Possibility Theory, Probability Theory and Multiple-valued Logics: A Clarification, *Annals of Mathematics and Artificial Intelligence* **32** 35–66.

[20] Dubois, D., Prade, H. (2008) An introduction to bipolar representations of information and preference. *International Journal of Intelligent Systems* **23** (8) 866–877.

[21] Dubois, D., Prade, H. (2006) A bipolar possibilistic representation of knowledge and preferences and its applications, *Fuzzy Logic and Applications* **3849** 1–10

[22] Dzeroski, S., Zenko, B. (2004) Is combining classifiers with stacking better than selecting the best one?. *Machine learning* **54** 225–273. Kluwer Academic Publishers.

[23] Fernández, A., Calderón, M., Berrenchea, E., Bustince, H., Herrera, F. (2010) Solving multi-class problems with linguistic fuzzy rule based classification systems based on pairwise learning and preference relations. *Fuzzy Sets and Systems* **161** (23) 3064–3080.

[24] Fernández, A., Carmona, C. J., Del Jesus, M. J., Herrera, F. (2016) A View on Fuzzy Systems for Big Data: Progress and Opportunities. *International Journal of Computational Intelligence Systems* **9** (1) 69–80.

[25] Fernández, A., Herrera, F., Cordón, O., Del Jesus, M. J., Marcelloni, F. (2019) Evolutionary Fuzzy Systems for Explainable Artificial Intelligence: Why, When, What for, and Where to?, in *IEEE Computational Intelligence Magazine* **14** (1) 69–81.

[26] Flores-Vidal, P. A., Gómez, D., Montero, J., Villarino, G.: Classifying segments in edge detection problems. (2017) 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Nanjing, 2017, 1–6.

[27] Flores-Vidal, P. A., Villarino, G, Gómez, D., Montero, J. (2019) Classifying segments in edge detection problems. International *Journal of Computational Intelligence Systems* **12** (1) 367–378.

[28] Franco, C., Montero, J., Rodríguez, J.T. (2012) A fuzzy and bipolar approach to preference modeling with application to need and desire. *Fuzzy Sets and Systems*, **214** 20–34.

[29] Friedman, J., Hastie, T., Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *Annals of Statistics* **28** (2) 337–407

[30] Friedman, H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics* **29**, 1189–1232.

[31] Fünkranz, J., Hüllermeier, E., Loza Mencía, E., Brinker, K. (2008) Multilabel classification via calibrated label ranking. *Machine Learning* **73** (2), 133–153.

[32] García S. and Herrera F., (2008) An extension on statistical comparisons of classifiers over multiple datasets for all pairwise comparisons, *Journal of Machine Learning Research* **9** 2677–2694.

[33] García S., Fernández A., Luengo J. and Herrera F., (2010) Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power, *Information Sciences* **180** (10) 2044–2064.

[34] García, V., Sánchez,J.S., Mollineda, R.A., (2012) On the effectiveness of preprocessing methods when dealing with different levels of class imbalance, *Knowledge Based Systems* **25** (1) 13–21.

[35] Gelman, A., Jakulin, A., Pittau, M. G., Su, Y. (2009). A Weakly Informative Default Prior Distribution For Logistic And Other Regression Models. The Annals of Applied Statistics, 2(4), 1360-1383.

[36] Gómez, D., Montero, J., Yanez, J. (2006) A coloring fuzzy graph approach for image classification, *Information Sciences*, **176** (24) 3645–3657.

[37] Gómez, D., Montero, J. (2004) A discussion on aggregation operators. *Kybernetika*, **40** (1), 107–120.

[38] Gómez, D., Rodríguez, J.T., Montero, J., Bustince, H., Barrenechea, E. (2016) n-Dimensional overlap functions. *Fuzzy Sets and Systems*, **287** 57–75.

[39] Goldberg D (1989) Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, Boston.

[40] Gupta, M. M., Zadeh, L. A. (1985) Approximate reasoning in expert systems, North-Holland, Amsterdam.

[41] Holland JH. (1975) Adaptation in Natural and Artificial Systems. The University of Michigan Press, Ann Arbor.

[42] Holm S., (1979) A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics* **6** 65–7.

[43] Hosmer, D.W., Lemeshow, S. (2000). Applied Logistic Regression. John Wiley and Sons.

[44] Hullermeier, E. (2005) Fuzzy methods in machine learning and data mining: Status and prospects, *Fuzzy Sets and Systems*, **156** (3) 387–406.

[45] Huhn, J., Hullermeier, E. (2009) FURIA: an algorithm for unordered fuzzy rule induction, *Data Mining and Knowledge Discovery*, **19** (3) 293–319.

[46] Ishibuchi, H., Nakashima, T., Nii, M., (2004) Classification and modeling with linguistic information granules: Advanced approaches to linguistic data mining.

[47] Kacprzyk J., Zadrożny S. (2016) Compound Bipolar Queries: A Step Towards an Enhanced Human Consistency and Human Friendliness.*Challenges in Computational Statistics and Data Mining. Studies in Computational Intelligence*, **605** 93–11.

[48] Kuhn M. (2008) Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, **28** (5) 1–26.

[49] Kumar, R. and Verma, R. (2012) Classification algorithms for data mining: A survey, International Journal of Innovations in Engineering and Technology, **2** 7–14.

[50] Kuncheva, L. (2000) Fuzzy classifier design. Studies in Fuzziness and Soft Computing. Heidelberg: Springer Verlag.

[51] Lim, TS., Loh, WY. and Shih, YS. (2000) A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms, Machine Learning, **40** 203–228.

[52] Lucasius C. and Kateman. G. (1993) Understanding and using genetic algo-rithms. Part 1. Concepts, properties and context. *Chemometrics and Intelligent Systems*, **10** 1–33.

[53] Lucasius C. and Kateman. G. (1994) Understanding and using genetic algorithms. Part 2. Representation, configuration and hybridization. *Chemometrics and Intelligent Laboratory Systems*, **25** 99–145.

[54] Magdalena, L. (2019) Semantic interpretability in hierarchical fuzzy systems: Creating semantically decouplable hierarchies. Information Sciences **496** 109–123.

[55] Magdalena L. (2015) Fuzzy Rule-Based Systems. In: Kacprzyk J., Pedrycz W. (eds) Springer Handbook of Computational Intelligence.

[56] Magdalena L., Gómez D., Montero J., Cubillo S., Torres C. (2019) Generalized Pre-aggregations. *Advances in Intelligent Systems and Computing*, **1000** 362–370.

[57] Mirjalili, S., Mirjalili, S. M., Lewis, A (2014) Grey Wolf Optimizer. *Advances in Engineering Software* **69** 46–61.

[58] Montero, J., Gómez, D., Bustince, H. (2007) On the relevance of some families of fuzzy Sets, *Fuzzy Sets and Systems*, **158** (22) 2429–2442

[59] Montero J., Bustince H., Franco C., Rodríguez J.T., Gómez D., Pagola M., Fernandez J., Barrenechea E. (2016) Paired structures in knowledge representation. *Knowledge-Based Systems*, **100** 50–58.

[60] Nauck, D. and Kruse, R. (1997) A neuro-fuzzy method to learn fuzzy classification rules from data, *Fuzzy Sets and Systems* **89** (3) 277–288.

[61] O'Doherty, J., Kringelback, M., Rolls, E., Hornak, J., Andrews C. (2001) Abstract reward and punishment representations in the human orbitofrontal cortex *Nat. Neurosci.*, **4** 95–102

[62] Ozturk,M., Tsoukiàs, A. (2007) Modeling uncertain positive and negative reasons in decision aiding. *Decision Support Systems*, **43** 1512–1526

[63] Pradera, A. (2008) Uninorms and Non-contradiction, *Modeling Decisions for Artificial Intelligence, Proceedings*, **5285** 50-61

[64] Ripley, B. D. (1996) Pattern Recognition and Neural Networks. Cambridge.

[65] Rodríguez, JT; Vitoriano, B; Montero, J. (2010) Una formulación axiomática de la noción de antagonismo semántico. Actas del III Simposio sobre Lógica Fuzzy y Soft Computing (LFSC2010) 181–188, Garceta, Madrid. ISBN: 978-84-92812-65-3

[66] Rodríguez, JT; Franco, C; Vitoriano, B; Montero, J. (2011) An axiomatic approach to the notion of semantic antagonism. Proceedings of the 2011 World Congress of the International Fuzzy Set Association (IFSA-AFSS), FT104-1/6, Indonesia, 2011. ISBN: 978-602-99359-0-5.

[67] Rodríguez, J.T., Turunen, E.,Ruan,D.,Montero, J. (2014) Another paraconsistent algebraic semantics for Lukasiewicz-Pavelka logic. *Fuzzy Sets and Systems.* **242**, 132–147.

[68] Rodríguez, J. T., Vitoriano, B., Montero, J. (2011) Rule-based classification by means of bipolar criteria. *IEEE Symposium on Computational Intelligence in Multicriteria Decision-Making (MDCM)*, 197–204.

[69] Rodríguez, J. T., Vitoriano, B., Montero, J. (2012) A general methodology for data-based rule building and its application to natural disaster management. *Computers & Operations Research*, **39** (4) 863–873.

[70] Rodríguez JT, Vitoriano B, Gómez D, Montero, J. (2013) Classification of Disasters and Emergencies under Bipolar Knowledge Representation. *Atlantis Computational Intelligence Systems* **7** 209–232.

[71] Rojas, K., Gómez, D., Montero, J., Rodríguez, J. T., Valdivia, A., Paiva, F. (2014) Development of child's home environment indexes based on consistent families of aggregation operators with prioritized hierarchical information. *Fuzzy sets and Systems*, **241** 41–6.

[72] Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, **1** 206–215.

[73] Sanz, J., Fernández, A., Bustince, H. & Herrera, F. (2013) IVTURS: A Linguistic Fuzzy Rule-Based Classification System Based On a New Interval-Valued Fuzzy Reasoning Method With Tuning and Rule Selection. *IEEE Transactions on Fuzzy Systems.* **21** 399–411.

[74] Septem Riza, L., Bergmeir, C., Herrera, F. and Benítez, J. (2015) frbs : Fuzzy Rule-Based Systems for Classification and Regression in R. *Journal of Statistical Software* **65** (6) 1–30.

[75] Triguero, I. et al. (2017) KEEL 3.0: An Open Source Software for Multi-Stage Analysis in Data Mining. *International Journal of Computational Intelligence Systems*, **10** 1238–1249.

[76] Trillas E., Moraga C., Guadarrama S., Cubillo S.,Castineira E. (2007) Computing with antonyms. Forging New Frontiers: Fuzzy Pioneers I, **217** 133–153.

[77] Turunen, E., Ozturk, M., Tsoukiàs, A. (2010) Paraconsistent semantics for Pavelka style fuzzy sentential logic. *Fuzzy Sets and Systems*. **161**, 1926–1940

[78] Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth edition. Springer.

[79] Villarino, G., Gómez, D., Cintas, R., Rodriguez, J.T. (2016) Metodología de minería de datos para el estudio de tablas de siniestralidad vial. In Proceedings of CAEPIA-STYLF Congress 2016 599–608 .

[80] Villarino, G., Gómez, D., Rodriguez, J. T. (2017) Improving Supervised Classification Algorithms by a Bipolar Knowledge Representation. Advances in Intelligent Systems and computing **643** 518–529.

[81] Villarino, G., Gómez, D., Rodriguez, J.T. et al. (2018) A bipolar knowledge representationmodel to improve supervised fuzzy classification algorithms *Soft Computing* **22** 5121–5146.

[82] Villarino, G., Gómez, D., Rodriguez, J.T. (2018) Assessing the performance of bipolar classifiers in three-class problems. In Proceedings of CAEPIA Congress 2018 306–314.

[83] Wilcoxon F. (1945) Individual comparisons by ranking methods, *Biometrics* **1** 80–83.

[84] Willighagen E. (2005) genalg: R Based Genetic Algorithm. `cran.r-project.org/`

[85] Wang, L., Mendel, J. (1992) Generating fuzzy rules by learning from examples, *IEEE Transactions on Systems, Man, and Cybernetics* 22 (6) 1414–1427.

[86] Yacubian, J., Gläscher, J., Schroeder, K., Sommer, T., Braus, D., Büchel, Ch. (2006) Dissociable systems for gain- and loss-related value predictions and errors of prediction in the human brain *Journal of Neuroscience* **26** 9530–9537.

[87] Zadeh, L. A. (1965) Fuzzy Sets, Information and Control, 8 (3) 338-353

[88] Zadeh, L. A. (1975a) Concept of a Linguistic Variable and Its Application to Approximate Reasoning.3., Information Sciences, **9** (1) 43–80

[89] Zadeh, L. A. (1975b) Concept of a Linguistic Variable and Its Application to Approximate Reasoning .1., Information Sciences, **8** (3) 199–249

[90] Zadeh, L. A. (1975c) Concept of a Linguistic Variable and Its Application to Approximate Reasoning .2., Information Sciences, **8** (4) 301–3574

[91] Zadeh, L. A. (1994) Soft Computing and Fuzzy-Logic, *IEEE Software* **11** (6) 48–56

[92] Zadeh, L. A. (2012) Computing with Words: Principal Concepts and Ideas. *Springer Publishing Company, Incorporated.* ISBN-3642274722 9783642274725.

[93] Md Zahangir Alom et al. (2019) A State-of-the-Art Survey on Deep Learning Theory and Architectures. *Electronics* **8** 292.